DOCUMENT RESUME

ED 145 829 IR 005 240

AUTHOR	Sager, Naomi
TITLE	Evaluation of Automated Natural Language Processing in the Further Development of Science Information Retrieval. String Program Reports No. 10.
INSTITUTION	New York Univ., N.Y. Linguistic String Project.
SPONS AGENCY	National Science Foundation, Washington, D.C. Div. of Science Information.
PUB DATE	Jul 76
GRANT	GN39879
NOTE	118p.
EDRS PRICE	MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS	Artificial Intelligence; *Automatic Indexing; *Computational Linguistics; Evaluation; Information Processing; *Information Retrieval; Information Systems; Language Classification; Man Machine Systems; *Science Materials; *Semantics
IDENTIFIERS	*Natural Language Processing

ABSTRACT

This investigation matches the emerging techniques in computerized natural language processing against emerging needs for such techniques in the information field to evaluate and extend such techniques for future applications and to establish a basis and direction for further research toward these goals. An overview describes developments in the information field which have led to renewed interest in natural language processing, sketches of programs for processing natural language to fulfill language-based functions of information systems, and the relationship between information and language. The stages of processing unrestricted natural language input of scientific communication into data structures suitable for information processing--parsing, structural transformations of parse outputs, and arriving at an underlying semantically meaningful representation--are outlined. The report also describes research related to the computerized discovery of semantic structures in science subfields; this research is concerned with the problem of structuring a data base which is given in natural language. Examples and suggestions for the application of techniques currently available or under development to information problems, and suggestions for further research in the language area of information science are presented. (Author/KP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: IR SE

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view. FL

ED145829

IR005240

String Program Reports No. 10

EVALUATION OF AUTOMATED NATURAL LANGUAGE PROCESSING IN THE FURTHER DEVELOPMENT OF SCIENCE INFORMATION RETRIEVAL

Naomi Sager

New York University

FINAL REPORT ON RESEARCH GRANT #GN39879

The Division of Science Information of the
National Science Foundation

NYU Linguistic String Project
251 Mercer St.
New York, New York 10012

July 1976

Contents

	Page
Preface	ii
1. Introduction	1
1.1 The Technological Setting	1
1.2 Tasks and Tools	5
1.3 Language Structure and Information	12
2. Natural Language Processing Techniques	15
2.1 Parsing	17
2.2 Transformations	25
2.3 Underlying Representations	32
3. Computerized Discovery of Semantic Structures in Science Subfields	45
4. Applications	72
4.1 Generating Word Lists	72
4.2 Generating Text Descriptors	75
4.3 Structuring Natural Language Data Bases	79
4.4 Aiding the User	88
5. Research Goals in Language and Information	93
References	101

PREFACE

In the course of making this investigation and writing this report I have become increasingly optimistic about the future of natural language processing in information retrieval and data processing. Over the past 18 years I have witnessed and participated in the development of natural language processing techniques which are now, finally, reaching the stage of useful application. This comes at a time when changes in the technology of information production and dissemination are making computerized natural language data bases widely available. Where computer readable natural language data bases exist, there soon develops a desire for computer programs to process the data in this form. In this setting, the natural language processing techniques which have been developed over the years take on a new importance and are challenged by the demand for new, natural language-based, information services.

The purpose of this investigation was to match the emerging techniques in computerized natural language processing against emerging needs for such techniques in the information field, to evaluate and extend such techniques as seemed most promising for future applications and to lay forth a basis and direction for further research toward these goals.

The organization of the report reflects the evolution of the investigation and the steps that led to the positive conclusion stated above.

Section 1 provides an overview of the background and results of the study. 1.1 summarizes developments in the information field which have led to a renewed interest in natural language processing. 1.2 sketches how programs for processing natural language could be used to fulfill some of the new language-based functions of information systems. 1.3 goes more deeply into the basis for such programs in the inherent relation between information and language, and traces some of the implications of this relation for a theory of information.

Section 2 describes in some detail the stages of processing which take largely unrestricted natural language input of the type encountered in scientific communication into the data structures suitable for information processing. These states are (roughly) parsing (2.1), performing certain structural transformations on parse outputs (2.2) and arriving at an underlying semantically meaningful representation (2.3). Some of the problems which were barriers to the success of most early programs have since been solved, and examples of solutions are described.

In the course of preparing section 2, it became clear that however broad or narrow the grammar required for a particular application, it was necessary in all cases to introduce constraints due to the specific subject matter, giving rise to a so-called semantic stage of processing. What was most troubling about this fact in relation to treating scientific discourse was the question of how the required semantic characterizations were to be arrived at. Even very restricted "worlds" are

difficult to specify a priori in all semantic detail. For fields of science, where such characterizations could be said to be the very goals of research, it is fruitless to build program structures which depend on having such characterizations supplied. How then overcome this new barrier? The research reported in Section 3, which was carried out largely as part of this investigation, is a partial answer to this question.

Moving along to applications, Section 4 presents some examples and some suggestions as to how techniques which are currently available or currently under development might be usefully applied to information problems. I have tried in section 4 to remain as concrete as possible, reporting only on work where some experimental results have been obtained.

Section 5 is a somewhat more relaxed attempt to project the same approach which was used in the more technical sections into suggestions for further research in the language area of information science. I would be grateful for critical comments on this section from readers who are also concerned with these problems.

Throughout this report, most of the examples are from work by myself and my research associates, referred to collectively as the Linguistic String Project (LSP). Aside from the obvious reason that I am most familiar with this work, the fact is that over the past 10 years there have been very few, if any, groups besides ourselves which have been concerned with natural language processing as applied to textual material, particularly the textual material of scientific discourse. I

hope that the results reported here, and the new demand for such research, will encourage others to join in this line of endeavor. While this report was written by myself alone, with the editorial assistance of Cynthia Insolio, it draws on the contributions of others who worked on this project, principally, Lynette Hirschman, Ralph Grishman, Carol Raze and Jerry Hobbs.

While preparing this report I have had several occasions to make oral presentations of materials from the investigation. The research results reported in Section 3 were presented at a Colloquium of the Division of Science Information of the National Science Foundation in November 1974 and at an Artificial Intelligence Symposium held at the Courant Institute of Mathematical Sciences of New York University in December, 1974. The long-range view expressed in Section 5 was presented and discussed at a workshop on Research Directions in Information Science in Anaheim, California in November, 1974. Parts of Sections 2 and 4 were utilized in preparing a perspective paper for the Workshop on Linguistics and Documentation of the International Federation of Documentation held in Stockholm in May, 1976. The Proceedings of this workshop will be published. The response to these presentations was most encouraging and I am grateful to the participants in the two workshops and the Symposia mentioned above for an exchange of information and points of view which aided in this study.

Naomi Sager

1. INTRODUCTION

1.1 THE TECHNOLOGICAL SETTING

That computerized natural language processing might be a useful tool in information retrieval was an idea that came early in the history of computers, and at a time when information retrieval was still a largely unfamiliar term. In the 1950's, administrators in the Office of Science Information Service* of the U.S. National Science Foundation were already considering this possibility. At that time the impetus was theoretical. The enormous power of the computer as a symbol processor was beginning to be appreciated. At the same time, the discovery in linguistics of transformations appeared to open the way for linguistic normalization of texts by formal, in principle, computable, methods. It seemed reasonable to project that computer programs for language analysis could be used to extract the topics and relations of interest in scientific documents by direct processing of the texts themselves. The basis for this clearly existed though the prospect was not immediate.

Over the next two decades, information retrieval systems grew very rapidly. The volume of scientific publication was ever-increasing, as were scientists' needs to keep up with this "information explosion." The information field could not wait for the results of long-term research. Information retrieval systems which processed computer readable forms of bibliographic data bases were built, and provided a valuable, if not perfect or complete,

*Now the Division of Science Information.

screening of the literature for the busy scientist. Among other measures of success was the tangible fact that individuals and organizations were willing to buy information retrieval services.

Because information retrieval succeeded to a large extent without natural language processing (NLP), some of the original impetus toward developing and incorporating NLP techniques into information systems diminished. Long-range research in NLP continued, but pressure from the information field for such techniques was minimal. Today this situation is changing. In recent years, a number of information scientists have pointed to limitations in present retrieval systems and stressed the need for more user-orientation, including more use of natural language. For example, in a 1971 panel [1] it was noted that there was a movement away from controlled vocabularies and toward natural language processing.* In more recent discussions, technological developments which enable the user to be in on-line contact with a variety of information resources, both of the bibliographic type and data banks, are leading to even greater emphasis on natural language processing. For example, in [2] in the section on information systems in the future, F. W. Lancaster writes:

> More and more data bases will exist in natural-language form and, as more efficient and economical

---

*In the context of the discussion in [1], "natural language processing" could be taken to include text scanning as well as more complex programs of the type referred to by the term "natural language processing" in this report.

> mass storage devices are developed, full text searching will become the norm in information retrieval operations. The pattern is more or less inevitable: more data bases in natural language form because "publication" itself will be electronic; more searching of data bases directly by scientists because these files will be readily accessible through terminals in offices and homes; more need for a natural language search approach because the person who is not an information specialist will not want to learn the idiosyncracies of a conventional controlled vocabulary and, even if he were willing to master one controlled vocabulary, the range of data bases that will be readily accessible to him virtually precludes the conventional controlled vocabulary approach.
>
> Those concerned with the design of information systems should now be concentrating on functional requirements for the user-oriented, natural language systems of the future.

There are several different ways in which natural language processing (NLP) may enter the information retrieval process in the future. For one, natural language would be the preferred medium of communication of the user with the system, as suggested by the above quotation. This implies that NLP techniques could handle the dialogue aspect of natural language communication and would also be able to sort, or filter, requests according to the type of response which would be appropriate. NLP techniques of this kind could serve as a front end to existing document retrieval systems as well as to systems which include other information services. NLP techniques could also be important in locating and retrieving specific information, i.e. in "fact retrieval," and by extension of some of the more sophisticated programs now being developed could, in restricted subject matter areas, perform some of the data processing

functions of research and clerical assistants, e.g. sorting, filing, retrieving, and summarizing information in natural language from various sources according to given categories.

The latter functions are, of course, a tall order for computers. However, the basis for such programs is being laid in the work described in 4.3, below, and the demand for such services is growing. The fact that large computerized stores of information in natural language form are being created in publishing and in record-keeping is stimulating a demand for techniques which can process data in natural language form. This is true not only in the area of official scientific publication, where a change in printing technology makes such data bases possible, but also in the area of file management, where institutions in medicine, industry, and government find it convenient to put large stores of natural language records into computer-readable form, primarily for convenience of access and storage.

An example of this process on a small scale in an institutional framework is the case of a hospital which computerizes its patient files for quick back-up to the written charts and for transactional purposes; then, finding itself with this large natural language data base, seeks computer techniques for processing the contents of the documents to obtain the summaries and other information required for health care evaluation. We are bound to see a pressure of this sort arising wherever natural language files are computerized. Once information is available

in computer readable form, users inevitably want the computer to process it.

1.2 TASKS AND TOOLS

The preceding section concluded with two main points:

(1) Natural language data bases are being created as part of the computerization of publication and file management;

(2) the existence of natural language data bases engenders a pressure to communicate in natural language with these information stores and to process them with a high degree of machine intelligence.

The question that immediately arises is: On what basis can we expect that computer programs will be developed that are capable of performing such tasks? And if there is indeed reason to believe that machines can perform these tasks, to what extent are the necessary techniques already in existence or being developed?

Regardless of the specific task, an advanced language capability requires that computer programs be able to recognize in natural language input the important content items and, to some level of specificity, the relations among these items so that a correct response results. Two features of natural language constitute a basis for such a capability. These features are particularly prominent in the language of science reporting and communication. One feature is the strong relation between

form and content; the natural language "code" for scientific information is so closely attuned to the content which it carries that the linguistic structure itself is part of the information. This point will be elaborated in the chapters which follow. The second feature is redundancy on the level of individual word occurrence and the occurrence of characteristic word-class combinations. Connected discourse requires word repetition [3]. Often, particularly in science writing, word repetitions are explicit, but there are many implicit word occurrences which are reconstructible on the basis of parallel syntactic structure and word-class cooccurrence patterns common in the text or in the domain of discourse. Reconstructions of implicit word occurrences and word cooccurrence patterns can be established objectively--in fact, computationally--within a science sublanguage, and to some extent in science writing as a whole. (This point is elaborated in relation to the analysis of conjunctional constructions in 2.2 below and in relation to science sublanguages in section 3.)

Computational methods based on these features of language, in particular as developed by the Linguistic String Project (LSP), make it possible to process science sentences into "stripped down" informational statements which are simpler and more regular than the initial sentences. The syntactic structure of the sentence provides a starting framework. But the syntactic structure is only in loose relation to the information content of the sentence due to the effect of linguistic

devices which enable the same content to be expressed in different grammatical forms. These devices are associated with transformations in linguistic analysis. A transformational stage of processing removes grammatical redundancies, fills in many of the implicit word occurrences, and identifies operators which have a fixed semantic effect. The result is a structuring of the sentence in which informational relations (e.g. a predicational hierarchy, if present) are explicit and grammatical redundancies are removed. At this level of analysis, sentences with the same information (up to word-synonymity) have the same representation. The steps of the computation and examples of the output at each stage of processing are given in section 2, below.

Such results are extremely encouraging with regard to the problem of establishing representations for natural language input which are suitable for information processing. Regular forms for the language input are essential if computers are to do the work of information processing. One hesitates to impose data structures on natural language input for the same reasons one avoids imposing multiple choice formats in many human communication tasks: The pre-set formats may not fit the content properly, they are on the whole unadaptable, and they may not be accepted by users. We are extremely fortunate, then, that the natural language form of science information lends itself to regularization out of features inherent in the language form itself. It turns out that several stages of processing are

required in order to bring out these regularities. Yet this can be considered a small price to pay in view of the new tasks that can be undertaken by natural language processors which are equipped to analyze and align the information carried in the natural language input.

It should not be assumed from the above discussion that the full power of the most advanced NLP techniques is required for every information task involving natural language, nor that NLP is required at all for many tasks. In the area of document retrieval, it is not at all clear that syntactic processing would increase recall and precision, at least not to the extent that warrants the investment in such processing. This is Gerard Salton's conclusion in his book Automatic Information Organization and Retrieval [4:p. 199], although it should be noted that Professor Salton's experiments were carried out using an early parser that may have lacked the discrimination necessary for a fair test of syntactic analysis. Karen Sparck Jones and Martin Kay in their review of the literature on this question [5] also find that "insofar as systematic comparative experiments have been carried out, they show that syntactic information contributes little to retrieval performance and may even detract from it" (p. 119). However, these authors suggest that the experiments may not have been adequate and the methods of handling syntax may not have been appropriate. It is generally agreed that the definitive experiments to evaluate the efficacy of syntactic analysis in retrieval have yet to be made.

Even in the absence of experiments that test NLP in conjunction with existing retrieval systems, experience with NLP systems would indicate that the best use, or at least the initial best use, of these powerful and expensive processors is in areas where their full power is needed, where no other techniques will do. Thus, one should look initially to applications of NLP in fact retrieval and fact processing, in query systems and on-line interactive programs rather than in document retrieval, as aids to improve performance. Later models of NLP systems, which have been perfected and streamlined in the course of fulfilling more restricted but necessary language tasks, may then be suitable and economical for use in document retrieval as well.

In the area of fact retrieval, several subtypes should be distinguished. Passage retrieval, developed by John O'Connor [6] returns answer-containing passages in response to factual queries. It achieves high precision and high recall without the use of a powerful natural language processor. The method is to expand the query into search word lists (a thesaurus-type operation) and then to match these lists against the words of the titles, abstracts, and figure captions of texts in the collection. It should be noted that in O'Connor's experiment, the computer search of figures in addition to titles and summaries increased recall from 35% to 80%. This illustrates the importance for fact retrieval of determining with some refinement where factual information of different types is located in science texts. It

also indicates that full text searching may not be required in order to achieve some of the effect of such searching.

A different type of fact retrieval is undertaken in question-answering systems. In their development to date, question-answering systems operate on structured data bases; this means that the "facts" are available from the outset in the form of a table. The processing consists of analyzing the user's factual query, which for the most part is in unrestricted natural language but is expected to pertain to the data base, translating the query into a representation which is consonant with the structure of the data base, matching the structured query with the data base and performing any auxiliary calculations needed to answer the query, and returning the appropriate reply to the user. These systems have a considerable history which has been reviewed periodically [7], [8]. In such well known systems as REL [9], REQUEST [10] and LSNLIS [11] the natural language aspect is well developed, at least with regard to interpreting questions within a narrowly circumscribed subject matter.

The linguistic problems in query analysis and interpretation do not seem to be different in kind from those faced in sentence analysis at the text end, with these differences: (1) A full grammar is not required; (2) Semantic constraints can be readily formulated and incorporated into the grammar because the domain is limited; (3) The data structures are given. These simplifications have enabled some system designers

to obtain a suitable transformation of the query for the given data base without solving the problem of transferability to other data bases. At the same time, the fact that the language problem was tractable has enabled these investigators to concentrate on the problems of matching the structured query with the data base and generating an answer. These are not trivial operations, because the correct response may require other operations on the data in addition to retrieval, and because in some cases implicit linguistic or logical links must be reconstructed. Question-answering is currently an active area of research and it seems likely that there will be a carry-over into practical information systems at the level of query interpretation and query/data matching.

Since the amount of scientific information which is available in tables independent of surrounding textual material is very small compared with that which is reported in natural language, a major focus of interest for the future is whether fact processing of the type practiced in advanced question-answering systems can be applied to information appearing in natural language. Work along these lines is in progress, starting with natural language records of a limited scope (e.g., medical records, weather reports) and applying linguistic programs that automatically transform the input into structures suitable for data processing. Here, the required NLP capability is much greater than is needed for query interpretation and the major design problem (ultimately a linguistic problem) is

whether procedures can be stated for regularizing textual material, where the procedures do not alter or lose any of the information in the text. This work is linked to a broader investigation into structural properties of the language of science.* The broader studies should indicate the type of data structures which are appropriate for characterizing and transfering science information, based on its language form. The computer results which have been obtained operating on narrow-domain texts demonstrate that automatic procedures can be applied to freely occurring natural language in a limited science area so as to produce data structures appropriate for very precise fact retrieval, fact comparison, statistical summary and the like [12]. While this work is in its infancy, the first results show that an NLP capability on this level of sophistication is not a speculation, but a real possibility.

1.3 LANGUAGE STRUCTURE AND INFORMATION

Information is not something separate from language. True, we convey some meanings by extralinguistic means. But for all practical purposes, and in most scientific and technical pursuits, the method of storing and transmitting information (and probably of forming new information) is largely linguistic. It is difficult to conceive of a theory of information that does not give some account of how information relates to language.

Because of the important connection between language and information, computational work on science texts takes on a theoretical as well as practical importance. In developing

*NSF Research Grant "Information Structures in the Language of Science," SIS75-22945, in progress.

computer techniques that begin with the natural language input and by successive stages of processing arrive at more regular forms for the information, we are both separating out the features of language which are not part of the information and arriving at a structural description of the information itself. As a result of following rigorous general methods, we have found that linguistically reported factual information has a simple overall structure. Stripped of stylistic variations, the information content of sentences can be represented by a hierarchy of predications plus conjunctions. This common frame of information can be viewed as a tree structure in which the nodes are words and the relation of predication is represented by a predicate word immediately dominating its argument words. Specific information is carried by the particular word values at the node positions. In science writing, the words at the different node levels fall into classes and there is a great deal of repetition and regularity in their distribution. This allows for stating much more detailed patterns that characterize the specific content of sentences over and above the overall predicational structure of the frame. (Examples are given in subsection 2.3 and section 3, below, and in [13] and [14].)

Much remains to be investigated along these lines. One question is the degree to which the detailed patterns mentioned above are common throughout science writing. Another has to do with the structure of argument and other discourse-level connections between sentences. It has been shown that

the methods of linguistic analysis can be extended in such a way as to isolate the elements of the line of argument in scientific papers [15]. However, as yet, we are further along on characterizing intra-sentence information structure than inter-sentence regularities.

2. NATURAL LANGUAGE PROCESSING TECHNIQUES*

A considerable literature has grown up concerning natural language processing (NLP) techniques. This report will not attempt to summarize this literature since several detailed review papers are already available. Covering broadly the whole area of computer programs for natural language are the reviews entitled "Automated Language Processing" in the Annual Review of Information Science and Technology, the two of current interest being the 1973 paper by Donald E. Walker [17], and the one by Fred J. Damerau in the forthcoming volume, to appear in October 1974 [18]. Both of these papers supply valuable summaries and extensive bibliographies of current work in the field, including topics which are outside of the scope of the present report, such as computer-based research in cognitive psychology, speech understanding, artificial intelligence, and computer-aided instruction. In addition, a survey of natural language parsing algorithms can be found in R. Grishman's Survey of Syntactic Analysis Procedures for Natural Language [19].

Two different contemporary approaches to the task of computer analysis of natural language can be distinguished. One approach, most commonly associated with research in artificial intelligence, is to build outward from "microworld" capability to larger domains; that is, to provide an integrated syntactic-semantic program which performs in a highly restricted environment,

*Material from this chapter and Chapter 4 appears in [16].

each successive program of this type being directed to a slightly less restricted domain. Following upon the initial work of Winograd [20] a growing literature of programs of this type exists (see, for example, recent issues of the American Journal of Computational Linguistics). It is not clear at this stage how far such programs will reach toward domains of scientific interest, such as one confronts in science information processing and information retrieval. However, some of the software developments in this research which concern general aspects of man-machine interaction may contribute to the design of user-driven front ends for information systems of the future.

The contrasting approach, which may be associated with more linguistically oriented NLP research, is to start from a broad linguistic capability, for example, a comprehensive computer grammar capable of analyzing arbitrary science text sentences, and to specialize and augment this capability in specific subject areas and for specific tasks. This approach can also be used to study the linguistic properties of subject-matter specialization in the language of science, both as a basis for establishing a language processing technology and as a theoretical tool to aid in understanding how science information is stored and transmitted via language. It is this second approach which is followed in most of the work described in this section.

In following this approach, we must be particularly concerned with methods that do not rely on an a priori statement of a complete semantics for the discourse that is to be

treated. This is extremely important in treating scientific communication where such material is unavailable, if not unattainable. Related to this criterion is that of extendability, since the effort involved in designing special programs for each new body of texts would be prohibitive. With these goals, successive stages of computerized analysis are seen to include parsing (2.1) and transformational decomposition (2.2), which are valid for the language as a whole, and the development of an underlying representation (2.3) which may be domain-specific. For science sentences, an underlying representation combines universal features of information representation (e.g., predication) with patterns of word-class co-occurrence which are special in each subject matter.

2.1 PARSING

Current use of the term "parsing" in the context of automatic language analysis refers to the assigning of a structure (or several alternative structures) to an input sentence. The structure is usually intended to represent the syntactic relations among the words of the sentence and often has the form of a tree diagram. The input sentence may in fact be fragments of a sentence, e.g. noun phrases, or texts consisting of several connected sentences and/or sentence fragments. It is permissible, however, to use the term "sentence" in this discussion to refer to any one of these input strings.

Despite differences in computational and linguistic frameworks it is generally agreed (in practice if not in

theory) that a syntactic analysis of a sentence is a necessary step in arriving at an explicit representation of its contents. While the gross structure of sentences in Indo-European languages is not difficult for us to state, its computation has not proved easy. For one reason, in the development of natural language parsers it was by and large assumed that syntactic structure meant phrase structure (e.g. structures obtained by such rules as: Sentence → Noun phrase + verb phrase, Noun phrase → Determiner + Noun, etc.), and that phrase structure analysis could be performed by context free parsers (recognizers limited to the above-type rules), at least to the extent that the resulting parse trees could be used as inputs for further processing. This effort was doomed to failure because in natural language, detailed constraints are involved at the various levels of description. Since the amount of detail that it is reasonable to express in context free rules is severely limited, this meant that the CF grammars that were used were necessarily inadequate. This problem is overcome by adding a component to the grammar which carries the detailed constraints, and applying these constraints to putative parse trees in the course of analysis. The parsing results then in fewer spurious analyses and the parsing time stays within reasonable (experimental) bounds.

Other problems in parsing, however, are related to the inherent complexity of natural language, to the existence of ellipsis, to the rich network of conjunction and comparative constructions, and to the syntactic ambiguities which remain even when a strong grammar with many detailed grammatical constraints

is used. The proper system framework is an important ingredient in the solution to some of these problems. It is a great help to work with a computational formalism that allows naturally for embedding, for applying constraints to parse trees, for dynamic generation of conjunctional constructions, for storing and passing along information conveniently, to mention just a few features which are now common in a number of systems. It should be kept in mind, however, that the use of a suitable formalism does not in itself provide solutions (a point which is sometimes overlooked by newcomers to the field). Where a creditable level of linguistic performance is required over a reasonably broad set of sentence types, there appears to be no substitute for an extensive grammar and word dictionary.

The problem of syntactic ambiguity is well illustrated by a passage quoted in Sparck Jones and Kay [5:p. 86] from Robinson and Marks [21] concerning a sentence for which they obtained 106 parses. The sentence was:

> An exploratory investigation is undertaken to survey the extent to which the idea of undetermined coefficients can profitably be used in practice for root-extraction with polynomial equations.

Robinson points out that "all sentences ending with two or more prepositional phrases or with a prepositional phrase following a noun object will receive multiple parsings; and if they may be attached to several verb forms, the multiplication of analyses is compounded further."

There is no general solution to this problem on the syntactic level, if one understands by "solution" the correct assigning of prepositional phrases (and other modifiers) to the element each modifies in the intended reading of the sentence. This would require applying selectional rules valid over the whole language, to use the distinction made in linguistics between grammatical rules and selection. Grammatical rules apply to all sentences of the language and deviations are recognizable as violations of grammaticality, whereas selection refers to the particular word-choice combinations that are made within grammatical categories to carry the specific content of the discourse. Certain word-choice combinations are frequent or accepted at a given time and in particular domains of discourse, but no hard and fast rules governing word-choice can be stated, as this would have the effect of legislating via the grammar what content can and cannot be said.

Since selectional constraints are in their nature subject-matter specific, they can only be applied when a particular subject matter, for which constraints are available, is being treated. Some workers have therefore limited their parsing to subject areas for which such rules can be stated <u>a priori</u>. Those engaged in less restricted natural language processing have developed means which are independent of subject-matter constraints in order to cope with--if not solve--the problem of multiple parses. For example, we can equip the parsing grammar with the ability to distinguish multiple readings of an associative kind

(e.g. the prepositional phrase problem), where the different readings do not affect the division of the sentence into its major syntactic components, from those which imply a radically different structural analysis of the sentence. We need not generate all the parse trees of the former type, provided subsequent stages of analysis can locate the implicated structure starting from any element in the sentence which might have a special relation to it. The Linguistic String Parser (LSP), for example, has such a pass-along device optionally available. The parser recognizes which structures are adjuncts (modifiers) and attaches a given adjunct, spanning a given sequence of sentence-words, in only one of the parse tree slots available for the given structure in otherwise identical parses. Later, when the transformational stage must examine some of these structures, chiefly prepositional phrases, to see if they are arguments of a nominalized verb--e.g. "of calcium" and "with other cations" as arguments of "exchange" in "The exchange of calcium with other cations"--a routine hunts for the prepositional arguments in all permissible adjunct slots of the nominalized verb. As semantic constraints are developed for the material, these can be applied in a similar manner.* The pass-along device enables the system to obtain correct syntactic structures without bogging down in excessive output at an early stage of analysis. Using this device, the LSP obtained one parse for the sentence discussed by Robinson and Marks, above (Fig. 2.1).

Reference has been made above to implicit word occurrences (ellipsis) and the importance of reconstructing these occurrences

*See the end of section 3.

Fig. 2.1

AN EXPLORATORY INVESTIGATION IS UNDERTAKEN TO SURVEY THE EXTENT TO WHICH THE IDEA OF UNDETERMINED COEFFICIENTS CAN PROFITABLY BE USED IN PRACTICE FOR ROOT-EXTRACTION WITH POLYNOMIAL EQUATIONS .

PARSE 1

SUCCESSFUL PARSE NO. 1
RUNNING TIME = 3.606
NODES IN TREE = 330
NODES ATTACHED = 1279

1. SENTENCE = INTRODUCER CENTER ENDMARK
2. .

2. ASSERTION = SA SUBJECT SA TENSE SA VERB SA OBJECT RV SA
3. INVESTIGATION IS 4.

3. LN = TPOS OPOS APOS NSPOS NPOS
AN EXPLORATORY

4. VENPASS = LVSA LVENR SA PASSORJ RV SA
UNDERTAKEN 5.

5. TOVO = LV &TO& VERB SA OBJECT RV SA
TO SURVEY 6. EXTENT 7.

6. LN = TPOS OPOS APOS NSPOS NPOS
THE

7. PWHS = P &WHICH& ASSERTION
TO WHICH 8.

8. ASSERTION = SA SUBJECT SA TENSE SA VERB SA OBJECT RV SA
9. IDEA 10. CAN PROFITABLY BE 11.

9. LN = TPOS OPOS APOS NSPOS NPOS
THE

10. PN = LP P NSTGO
OF 12. COEFFICIENTS

11. VENPASS = LVSA LVENR SA PASSOBJ RV SA
USED 13.

12. LN = TPOS OPOS APOS NSPOS NPOS
UNDETERMINED

13. PN = LP P NSTGO
IN PRACTICE 14.

14. PN = LP P NSTGO
FOR ROOT-EXTRACTION 15.

15. PN = LP P NSTGO
WITH 16. EQUATIONS

16. LN = TPOS OPOS APOS NSPOS NPOS
POLYNOMIAL

for purposes of text regularization. The problem begins at the level of parsing because of the necessity of handling sentences containing coordinate conjunctions and comparatives, where ellipsis is rife. For example, given a typical science sentence, such as "Infants and children absorb and excrete digoxin in much the same way as adults do" we may want for informational purposes the separate units:

(1) Infants absorb digoxin in much the same way that adults absorb digoxin,

(2) Children absorb digoxin in much the same way that adults absorb digoxin,

(3) Infants excrete digoxin in much the same way that adults excrete digoxin,

(4) Children excrete digoxin in much the same way that adults excrete digoxin.

The computational problem can be divided into three parts:* generating parse trees which cover the occurrences of conjunction-headed strings in sentences; locating words in the sentences which are repeated implicitly in a "zeroed," i.e. elided, form in particular positions in conjunction strings; and reconstructing the complete sentences underlying conjunctional occurrences when the application requires such an expansion. A solution to the first problem has been in effect in the LSP

---

*The material in this paragraph and the following two paragraphs is also presented, in a slightly different form, in [22].

system since its first implementation [23-25], and similar devices have been described more recently, e.g., by Woods in the framework of augmented transition networks [26]. A computational solution to the second problem, that of locating the words that are zeroed under conjunction, has been provided by Raze for the LSP system [22]. In this solution, which is based on general properties of conjunctional constructions, a mechanism locates zeroed elements in the conjunction strings and cross-references them with respect to related elements in the head construction. Constraints can then be applied to elided conjuncts as though they were expanded, and transformational expansion (the third problem, above) can then be carried out straightforwardly by following the pointers which have been set up.

The main innovation in Raze's algorithm is called "stacking." It is a nondeterministic programming device which causes restrictions (i.e., subprograms applying detailed constraints) to be re-executed on conjoined segments whenever the restriction is invoked on an element which has a conjunct. Since the constraint usually applies to a combination of words in the main string and in the conjunct, the device must ensure that the proper grammatical relation holds between these words. That is, it treats these separated words as a single linguistic entity, obtaining the effect of a full expansion of the conjunctional occurrence without carrying out the physical rearrangement and copying of the parse tree. The strategy of treating ellipsis in two steps (locating deleted elements and later carrying out the physical expansion) is important for a solution to conjunctions

for several reasons. First, it is costly and often fruitless to expand a conjunctional string until one is sure one has a good parse. Second, to get a good parse one has to execute restrictions on the generated parse tree including the conjunction subtrees, and this requires locating deleted elements. And third, the decision as to whether conjunctional constructions should be expanded or not is application-specific.

Raze's program appears to be the first implemented general algorithm to treat conjunctional ellipsis on a broad scale and in great grammatical detail. It is followed (in the LSP system) by a transformational expansion of conjunction and comparative strings, which was also programmed by Raze, and is illustrated below.

2.2 TRANSFORMATIONS

Parsing is rarely seen as an end in itself and often as the first step in a transformational decomposition. The great interest of transformations for linguistic computations is that their use enables the processor to reduce the number of alternative grammatical forms for the same information. Thus, for example, two sentences which contain, respectively, "Ca^{++} exchanges with other cations in SR" and "the exchange of Ca^{++} with other cations in SR," over these portions carry the same information, and it would be a definite gain to have only one representation for it. Furthermore, it is hoped that the structural representation of the sentence after a transformational decomposition is performed will be closer to an "ideal"

representation of its contents than either the original word string or, in most cases, the syntactic parse.

The main steps in computerized transformational analysis are first to obtain a parse or "surface structure" for the sentence, and then to apply the appropriate sequence of "reverse" transformations so as to arrive at the desired underlying representation or "deep structure".* This has to be repeated for each initial surface parse tree, and for all the alternative intermediate trees created when more than one reverse transformation applies to the tree resulting from the previous application of a reverse transformation. Clearly, the process explodes unless (a) the parser delivers a reasonably small number of alternative surface parse trees to begin with; and (b) the number of reverse transformations which can apply to any intermediate tree is severely constrained.

Early transformational programs foundered on these two counts. Parsing grammars, as we have seen, were not strong enough to deliver only correct parses, and the problem of ambiguity remained. With regard to the reverse transformations, most workers were using the transformational generative grammar (TGG) of Chomsky, which does not provide rules for going from the surface structure to the deep structure, and they did not have available detailed statements of the parse tree conditions

---

*The terms "surface structure" and "deep structure" are used here as they have been used in computational linguistics rather than as they are used in Chomsky's theory of transformational generative grammar.

which permit a given transformation to apply. The result was that after valiant efforts some transformational projects were abandoned; others limited their goals and attempted to supply the needed surface grammar and transformational pinpointing themselves. Later arrivals to computational linguistics sought alternative grammatical frameworks or turned away from grammatical analysis to semantics altogether.

Conditions (a) and (b) are somewhat easier to meet using the original formulation and subsequent development of transformations by Harris [27]. This theory provides an explicit statement of the changes in the sentence produced by each transformation. These changes have their counterpart in the surface parse tree. Given, then, that reasonable parses can be obtained by enriching the parsing grammar, and that the ambiguity problem can be contained by such devices as the pass-along mechanism described above, it is then a straightforward operation to test the parse tree for traces of a particular transformation in order to determine that the corresponding reverse transformation applies. In point of fact, using this approach, it is possible to build most of the information needed to select the correct reverse transformation into the node labels and node attributes of the surface parse tree itself. This makes many reverse transformations exceedingly simple, e.g. relative clauses, it-permutations, object sentence-nominalizations, and even conjunctional expansions, as has been the LSP experience.

Pairs of LSP outputs showing the surface parse followed by the transformational decomposition are shown in Figs. 2.2a, b

Fig. 2.2a

THE INHIBITION OF THE ATPASE PRODUCES THIS CHANGE IN AFFINITY .

PARSE 1

1. SENTENCE = INTRODUCER CENTER ENDMARK
2. .

2. ASSERTION = SA SUBJECT SA TENSE SA VERB SA OBJECT
3. INHIBITION 4. PRODUCES 5. CHANGE

3. LN = TPOS QPOS APOS NSPOS NPOS
THE

4. PN = LP P NSTGO
OF 7. ATPASE

5. LN = TPOS QPOS APOS NSPOS NPOS
THIS

6. PN = LP P NSTGO
IN AFFINITY

7. LN = TPOS QPOS APOS NSPOS NPOS
THE

Fig. 2.2b

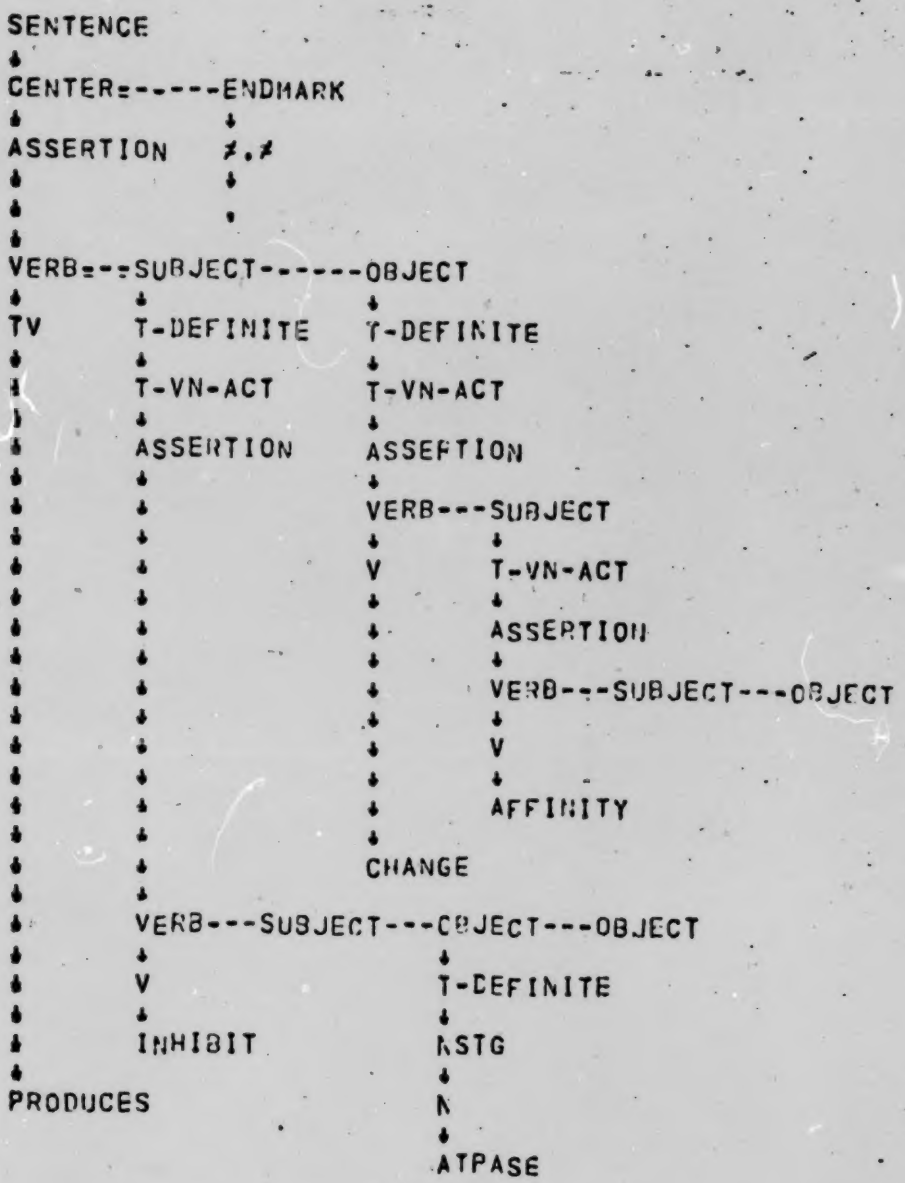

SENTENCE
CENTER
ENDMARK
ASSERTION
≠.≠
VERB
SUBJECT
OBJECT
TV
T-DEFINITE
T-DEFINITE
T-VN-ACT
T-VN-ACT
ASSERTION
ASSERTION
VERB
SUBJECT
V
T-VN-ACT
ASSERTION
VERB
SUBJECT
OBJECT
V
AFFINITY
CHANGE
VERB
SUBJECT
OBJECT
OBJECT
V
T-DEFINITE
INHIBIT
RSTG
PRODUCES
N
ATPASE

Fig. 2.2c

```
* BROSS  1. 1.18
THE HEART , LUNGS , AND BONY STRUCTURES
ARE INTACT .

   PARSE  1

1. SENTENCE  = TEXTLET
               2.

2. OLD-SENTEN= INTRODUCER  CENTER  ENDMARK
                           3.      .

3. ASSERTION = SA  SUBJECT          SA  TENSE  SA  VERB  SA  OBJECT   RV  SA
                   4. HEART , 5,                    ARE      INTACT

4. LN        = TPOS  OPOS  APOS  NSPOS  NPOS
               THE

5. O-CONJ    = NVAR       ANDSTG
               LUNGS ,    AND 6,

6. O-CONJ    = LN   NVAR
               7.   STRUCTURES

7. LN        = TPOS  OPOS  APOS  NSPOS  NPOS
                           BONY


TRANSFORMATIONS IN SENTENCE WHICH SUBSUMES THE HEART , LUNGS , AND BONY STRUCTURES
          ARE INTACT . :  TSEQ1 *

TRANSFORMATIONS IN TEXTLET WHICH SUBSUMES THE HEART , LUNGS , AND BONY STRUCTURES ARE
          INTACT . :  TSEQ2 *

TRANSFORMATIONS IN OLD-SENTENCE WHICH SUBSUMES THE HEART , LUNGS , AND BONY
          STRUCTURES ARE INTACT . :  TSEQ4 *

TRANSFORMATIONS IN CENTER WHICH SUBSUMES THE HEART , LUNGS , AND BONY STRUCTURES ARE
          INTACT :  TSEQ5 *

TRANSFORMATIONS IN ASSERTION WHICH SUBSUMES THE HEART , LUNGS , AND BONY STRUCTURES
          ARE INTACT :

T-EXPAND-TO-CENTER EXECUTED IN ASSERTION WHICH SUBSUMES THE HEART , LUNGS , AND BONY
          STRUCTURES ARE INTACT
    ( EXPAND -> ASSERTION=THE HEART , LUNGS , AND BONY STRUCTURES ARE INTACT ) *
          IMPLY( ( CONJUNCT ) * ) *
```


T-MOVE-RC * T-MOVE-NSTG-FRAG * T-FORMAT *

Fig. 2.2d

```
TRANSFORMATIONS IN LNR WHICH SUBSUMES THE HEART , LUNGS , AND BONY STRUCTURES (

T-EXPAND-TO-CENTER EXECUTED IN LNR WHICH SUBSUMES THE HEART , LUNGS , AND BONY
                STRUCTURES
    ( EXPAND -> LNR=THE HEART  ) + IMPLY( ( CONJUNCT -> LNR=THE LUNGS ,  ) + ITER(
            ( IMMEDIATE-NODE -> NSTG=THE HEART , THE LUNGS , AND BONY STRUCTURES
            ) + ( EXPAND -> NSTG=THE HEART  ) + IS(  ( CENTER )  ) - (
        IMMEDIATE-NODE -> SUBJECT=THE HEART , THE LUNGS , AND BONY STRUCTURES  )
         + ( EXPAND -> SUBJECT=THE HEART  ) + IS(  ( CENTER )  ) - (
        IMMEDIATE-NODE -> ASSERTION=THE HEART , THE LUNGS , AND BONY STRUCTURES ARE
         INTACT  ) + ( EXPAND -> ASSERTION=THE HEART ARE INTACT  ) + IS(  (
        CENTER )  ) - ( IMMEDIATE-NODE -> CENTER=THE HEART ARE INTACT , THE LUNGS
        , ARE INTACT AND BONY STRUCTURES ARE INTACT  ) + ( EXPAND -> CENTER=THE
        HEART ARE INTACT  ) + IS(  ( CENTER )  ) +  ) +  ) +

1. SENTENCE  = TEXTLET
               2.


2. OLD-SENTEN= INTRODUCER  CENTER  COMMASTG  ENDMARK
                           3.      , 4.      .


3. ASSERTION = SA  SUBJECT    SA  TENSE  SA  VERB  SA  OBJECT   RV  SA
                   5. HEART                   ARE      INTACT


4. Q-CONJ    = CENTER  ANDSTG
               6.      AND 7.


5. LN        = TPOS  QPOS  APOS  NSPOS  NPOS
               THE


6. ASSERTION = SA  SUBJECT    SA  TENSE  SA  VERB  SA  OBJECT   RV  SA
                   8. LUNGS ,                 ARE      INTACT


7. Q-CONJ    = CENTER
               9.


8. LN        = TPOS  QPOS  APOS  NSPOS  NPOS


               THE


9. ASSERTION = SA  SUBJECT          SA  TENSE  SA  VERB  SA  OBJECT   RV  SA
                   10. STRUCTURES                 ARE      INTACT


10. LN       = TPOS  QPOS  APOS   NSPOS  NPOS
                           BONY
```

and 2.2c, d. The pair 2.2 a, b illustrates how nominalized verb forms (inhibition, change) are restored to operator (verb) status with their proper arguments. The pair 2.2c, d shows the operation of conjunction-expansion on a sentence from a corpus of medical narrative which has been computer-analyzed by the LSP. Figure 2.2c shows the surface parse (string analysis) which is passed to the transformational component, and the trace of the first few steps of the transformational decomposition. Figure 2.2d (a later portion of the same trace) shows the conjunction-expansion transformation being executed and the resulting expanded decomposition, which now contains three similar ASSERTIONS each with one noun as subject, in place of one ASSERTION with three conjoined nouns as subject. Conjunction expansion is a crucial ingredient of text regularization, though the decision as to which types of conjuncts should be expanded and up to what level is sometimes application-specific. The code for the expansion transformation shown in Fig. 2.2d is written at an extremely high level of generality (partly indicated by the brevity of the trace) and operates effectively on a wide range of conjunctional constructions, virtually all those encountered in most scientific writing. A similarly general comparative-transformation has been written in the LSP framework and one also exists in the IBM REQUEST system. Both groups have a number of other transformations in stock or in process.

One should not conclude from the above that all the problems in automatic transformational analysis have been solved.

Transformational decomposition on a broad scale over arbitrary texts is still in the future. What has been achieved in the work illustrated above and by others (notably Petrick and colleagues at IBM) is a major step toward the goal. Also, the techniques that have been developed and the set of transformations which have been programmed so far do a useful amount of sentence regularization in themselves. Furthermore, because of their general linguistic basis, they should be applicable in widely different subject matter areas. Since the writing and debugging of transformations is a lengthy process, it is fortunate that a number of applications are possible without a complete transformational analysis (section 4 below). It is also possible to proceed with the expansion of the transformational component and with applications while problems on the discourse level (e.g. reference resolution and intersentence connectives) are being studied.

2.3 UNDERLYING REPRESENTATIONS

As noted earlier, the main motivation for using transformations in computational linguistics was the belief that such analysis would strip the sentence of its grammatical redundancies and provide a structure which was in closer harmony with the content of the sentence. Now that a considerable amount of work has been done we can ask to what extent this is the case and whether (or what kind of) further analysis is needed. The answer to this question clearly depends on the particular version of "deep structure" that is being considered and also on the

particular view that is held of sentence content. In practice, an overriding consideration is what one intends to do with the analysis.

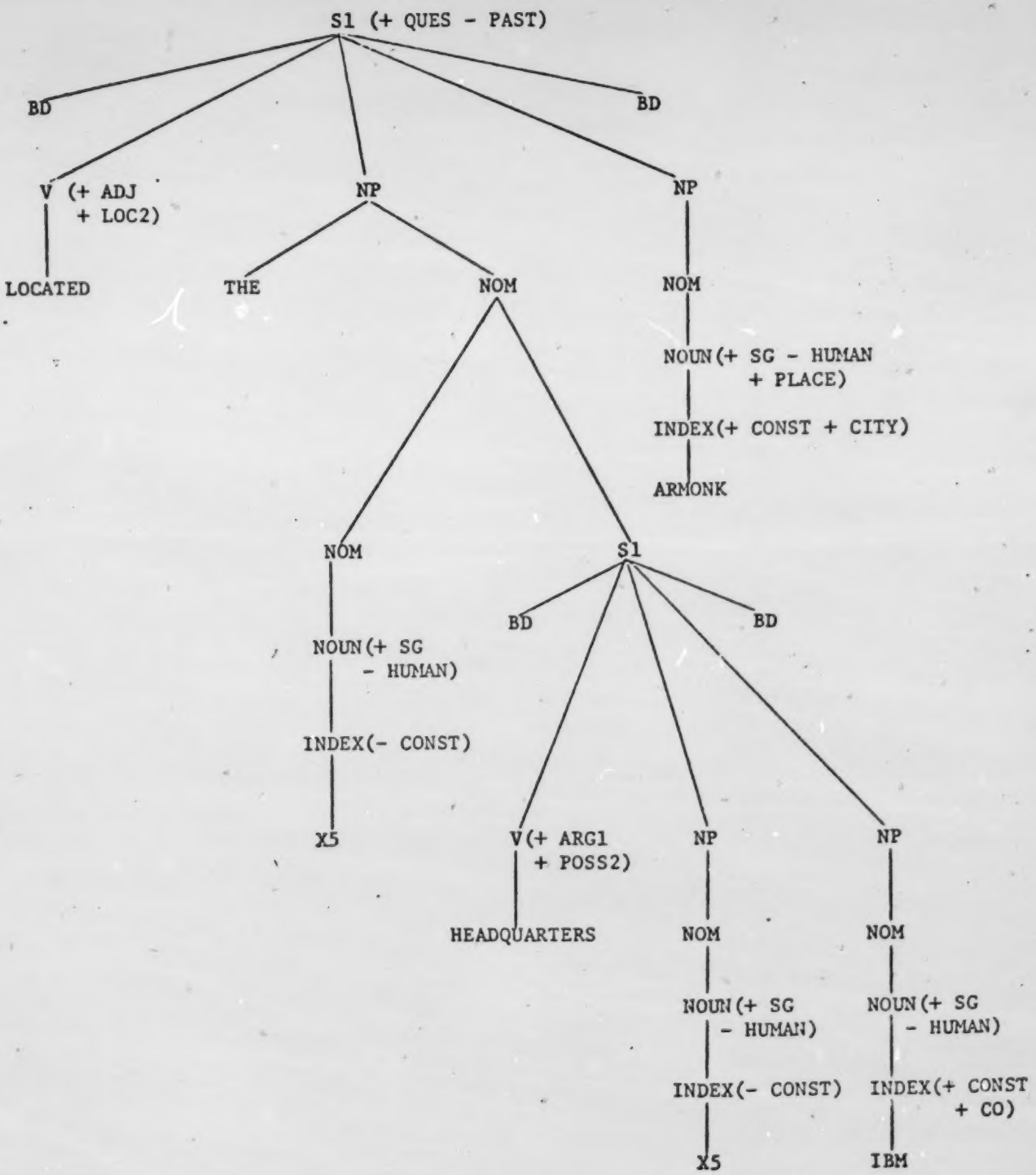
With regard to the different versions of deep structure, I intend here to bypass what most linguists (and many others) would consider the critical issue: the choice of the proper underlying structure. The reason for this, crudely stated, is that I do not believe this is an area in which we have much freedom of choice. Rather, we are given language data and we try to develop methods for treating these data; the structures we arrive at are the ones which result from applying the methods, and which can be shown to exist in a broad coverage of the language data. This is, of course, a simplification of the problem since we have some choice over the methods of analysis and these are intertwined with the theory of language structure which is used. Nevertheless, I believe this a relevant computational view, at least where applications are intended. The diverse structures which have been proposed for representing the underlying content of sentences become appropriate for consideration in an applicational context to the extent that they are accompanied by methods (including here linguistic content) which can be shown to apply to a reasonable corpus of non-contrived sentences. On this criterion, many of the interesting semantic theories that have been proposed are not at a stage where they can be considered as models for applications.

With regard to the two major transformational theories, transformational generative grammar, as incorporated into the

IBM REQUEST system (a system for answering queries from a structured data base of business statistics), produces deep structures of the type illustrated in Fig. 2.3. The underlying structures are trees consisting of one or more nested propositions (S1) each marked off by a pair of boundary symbols (BD). Each proposition consists of a predicate (V) followed by its associated arguments (NP's), which always occur in a fixed order in the structure. Plath [28] points out that such surface structure elements as auxiliaries, prepositions, inflectional endings, and punctuations--all major sources of syntactic variation--have been eliminated in favor of binary syntactic features (e.g., (-PAST), (+POSS2), (+SG), and (+QUES)) occurring at standard positions in the tree. He goes on to point out that a key property of these underlying structures is their close resemblance to expressions in the predicate calculus, e.g. as illustrated in Fig. 2.3, the representation of surface proper nouns ("IBM", "Armonk") as logical constants (INDEX (+CONST)) and the treatment of surface common nouns such as "headquarters" as variables (INDEX (-CONST)) in propositional functions. Following the transformational stage, the REQUEST system translates structures of this type into an underlying representation in "logical form," which is useful in the question-answering operation proper. Woods' LSNLIS system for answering questions from a structured data base concerning lunar rocks [11] also mapped into an underlying representation based on the predicate calculus.

In this connection it is worth noting that currently interest is growing in the data base management field in

Fig. 2.3



Underlying Structure of "Is IBM's headquarters in Armonk?"

From W. Plath, The Request System, IBM RC 5604, Yorktown Hts, N.Y. 1975.

alternatives to the predicate calculus as a representation for structured data bases in certain applications. In particular, the relational model of data introduced by E. F. Codd has stimulated the development of at least one data base query language (SQUARE), which the authors suggest "mimics how people use relations or tables to obtain information" and does not require the mathematical machinery of the predicate calculus in order to express simple references to tables.* Natural language processing front ends have been suggested for these systems. These developments are of interest both to computational linguistics and information science.

The use of Harris' transformational theory [27], as adapted for text processing by the LSP, results in sentence decompositions of the type illustrated in Fig. 2.4.** In Fig. 2.4, the form is a dependency tree where each node corresponds to a word or base-form of a word; if the word is a base form of a verb, or is by another criterion an "operator" in the sense of this theory, its arguments are shown as daughter nodes. For some purposes a linearization of the tree is desired; this is also shown.

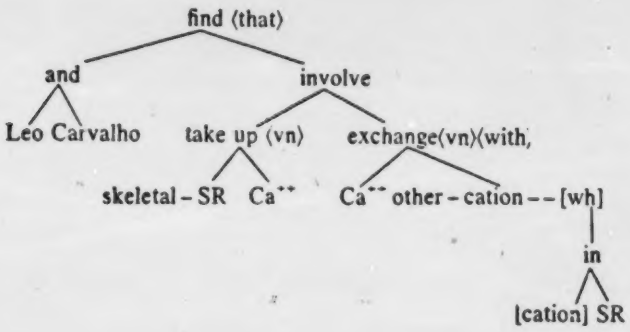
---

*References to Codd's papers are included in Refs. [29] and [30].

**The transformational decomposition operates on the string parse tree and converts it to a tree which is composed of simple ASSERTIONS connected via nodes specifying the transformations which operated [31]. A simple program converts outputs of this type to the type shown in Fig. 2.4.

Fig. 2.4

Carvalho and Leo found that the Ca^{++} uptake of skeletal SR involves the exchange of Ca^{++} with other cations in SR. (LE711 13C.5.7)



Linearization of example

(find ⟨that⟩
(and (Leo) (Carvalho))
(involve
(take up ⟨vn⟩
(SR-(skeletal)) (Ca^{++}))
(exchange ⟨vn⟩ ⟨with⟩
(Ca^{++}) (cation-(other) – – ([wh](in([cation])(SR)))))).

This sentence tree has several interesting properties from an informational point of view. First of all, it suggests that information, at least scientific information as it is carried by language, can be resolved into a hierarchy of predications. In the example sentence, what is predicated about Ca^{++} and other cations is that they exchange with each other, and what is predicated about skeletal SR (sarcoplasmic reticulum) and Ca^{++} is that the one takes up the other. This is from reading the bottom-most operators with their operands. Next, about the taking up and the exchange, it is predicated that the one involves the other. On the highest level of the hierarchy it is predicated that there is a relation between the human investigators Leo and Carvalho and the hierarchy of relations dominated by "involve," and that this relation is "find."

When this type of analysis is made over several hundred science sentences it can further be seen that certain classes of words characteristically are found at the bottom of the hierarchy and others near the top (when both classes are present in the sentence). The concrete words referring to objects in the science are at the bottom, and the intellectual verbs, like find, hypothesize, discover, which characteristically take human subjects, appear at the top; thus, the analysis effectively separates the "object language" of the science from the "meta language"--that is, it separates the words referring to the objects and relations in the science proper from statements the human investigator makes about those objects and relations. On

levels in between, one finds definite classes, like quantity operators and causality operators, each in characteristic positions vis à vis other operators. These classes and their positions in the hierarchy appear to be common in much of scientific writing. All of this is highly suggestive of a structure to the information which is being transmitted--a structure not identical with that of the sentences, but certainly in close relation to the linguistic frame.

Can this "deep structure" itself serve as the ultimate representation of sentences? For certain statistical processes, such as clustering to obtain semantic classes, this representation yields good results without further transforming. Thus, a clustering program described in section 3 and [32] successfully obtained a promising set of semantic classes for a subfield of pharmacology when it was applied to a sample corpus of linearized transformative trees of the type shown in Fig. 2.4.

For other purposes, such as question-answering, fact retrieval, and data processing, this representation, while undoubtedly capturing the overall fabric of the relations involved, lacks the detailed classification of operators and objects which would permit specific pinpointing of content or the retrieval of particular subsets across the hierarchies. It also lacks the simplicity and the logical structure which is sought in most data base applications.

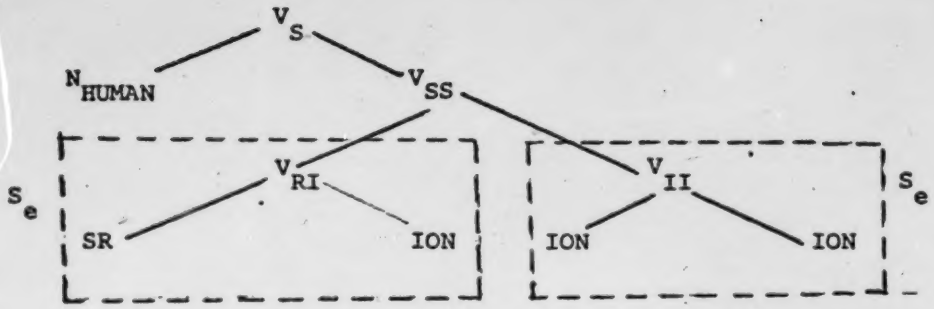
One way of arriving at structures that reflect specific textual content is to make use of the restrictions in language usage which are characteristic of the texts in a particular

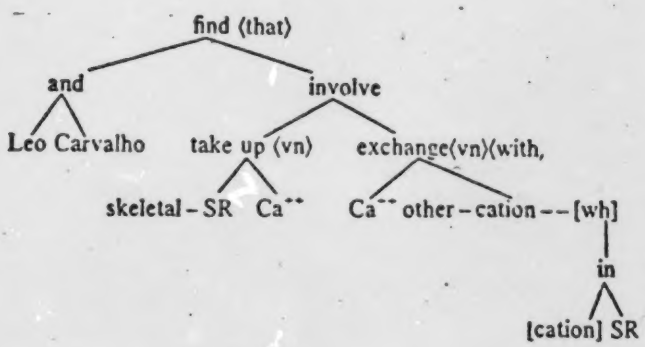
subject matter: that is, to exploit the fact that on a particular topic, only certain words in certain combinations actually appear. Distributional linguistic techniques applied to transformationally analyzed texts in a given subfield yield patterns of word-class cooccurrence. These patterns can be seen as selectional specializations of the transformational dependency trees. Instead of seeing the words at the nodes as members of the large grammatical classes "noun," "verb," etc., we can view them as members of the subfield classes (e.g. in the pharmacology data: the drug class, the ion class, etc.). We find that only certain combinations of word-class nodes occur. Each pattern can be represented by a transformational tree whose nodes are relabeled in accord with the subfield word classes which co-occur in that particular configuration. The interesting point is that a small set of patterns (relabeled trees) for a body of texts reflects specifically what is being talked about most in the texts. The relabeled trees are thus a type of format for the information in a great many sentences and sentence-portions dealing with the same subject matter.

To illustrate, a sentence format of the type found to be characteristic of the pharmacology subfield referred to above is shown in Fig. 2.5; along with a repeat of the sentence of Fig. 2.4, which is an instance of this format. The lower level subformats S_e cover the occurrences of elementary (object language) sentences. In the pharmacology material, two of these were of the types SR V_{RI} ION, ION V_{II} ION, appearing in Fig. 2.5

Fig. 2.5

Subfield format in tree form





V_S : verb class with sentential object, human subject (e.g. find, think assume)

V_{SS}: verb class with sentential subject and object, often causal in nature (e.g. influence, affect)

S_e : elementary sentence format, specific to the subfield; here involving:

V_{RI}: verb class connecting nouns of the SR class with those of the ION class (e.g. take up, contain)

V_{II}: verb class connecting nouns of the ION class to nouns of the ION class (e.g. exchange, compete)

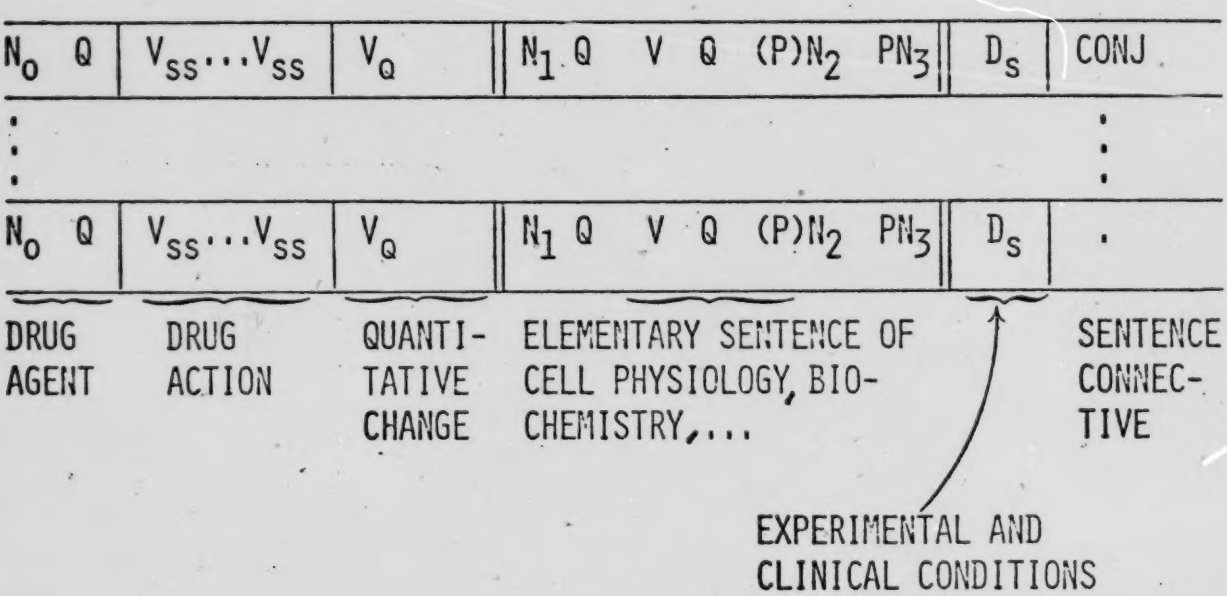
under V_{SS}. Of course, there are fairly numerous bottom level subformats for a subfield, and they appear in different combinations under V_{SS} (causal) verbs. This reflects the fact that scientific reporting is concerned with establishing causal connections between events. A feature not illustrated in Fig. 2.5 is the fact that, in this science at least, quantitative operators play a significant role. These characteristically occur between the elementary sentence level and the causal (V_{SS}) level. In contrast, time operators did not play a role in this material.

A summary of the features of the pharmacology subfield formats for the factual portions of sentences (i.e. cut off at V_S) obtained for the large bulk of the sentences in a 300-page corpus is shown in Fig. 2.6. Here, to illustrate the repetition of format occurrences in texts, each subformat S_e with its immediate operators up to the introduction of a drug word, if present, is shown "flattened," filling one line. Connectives between format lines may be conjunctions, but alternatively V_{SS} verbs as in Fig. 2.5. In a text occurrence, the innermost section (between double bars) contains a verb with its concrete noun subject and object(s), i.e. a bottom level operator in the transformational tree, along with its arguments (e.g. "Ca^{++} exchanges with other cations"); and the successive boxes to the left represent the hierarchy of verbal operators on it.

Linguistically based subfield formats are one answer to the question of underlying representation. While they are based on selectional constraints that operate in particular science

Fig. 2.6

PARTIAL TEXT FORMAT FOR
PHARMACOLOGICAL SUBFIELD



From N. Sager, The Sublanguage Technique in Science Information Processing, Journal of the American Society for Information Science, 26, 10-16, 1975.

subfields, they have certain features which may be common to many science fields. Just as one does not arrive at a universal grammar by studying all languages at once, it may be that in attempting to represent the semantic content of texts, we will be obliged to start with the regularities that obtain in demarcated domains of discourse. At the same time, obtaining the special forms which can be established for one subfield at a time is certainly not the only way to utilize the informational connections which emerge from transformational analysis. We may also consider each triple consisting of an operator with its two arguments (the most frequent case) as a unit. The operator, usually a verb, can be viewed as filling the role of F in the functional notation F(X)=Y.* X and Y may be concrete nouns, or themselves "F"s with their own arguments. The words which fill an "F" role on similar argument types can be grouped together and assigned a semantic label. Irwin Bross and associates at Roswell Park Memorial Hospital used such a final representation for linguistically analyzed surgical reports.** This may also be a useful form for retrieval, or for arriving at coordinated index terms from deep structures, as suggested by Sparck Jones and Kay [5] and treated in 4.2 below.

---

*For the moment, in this adaptation of transformations, we do not count among the F operators the connectives concerned with adjunction and conjunction because the deformed sentences under such connectives can be expanded to full subtrees, and the connection established by and, wh- is too general for practical use.

**Ref. [33] and other references noted therein.

3. COMPUTERIZED DISCOVERY OF SEMANTIC STRUCTURES IN SCIENCE SUBFIELDS*

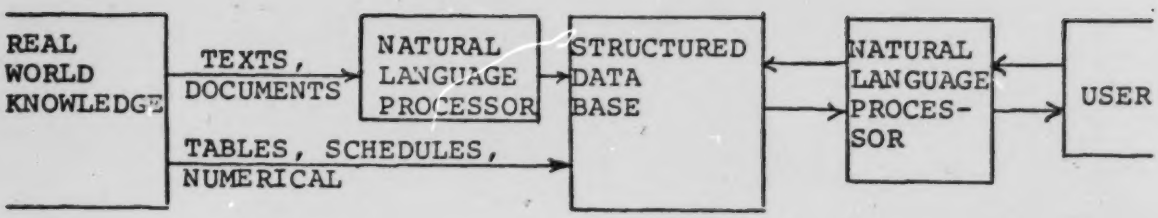
To bridge the gap between basic NLP research and informational applications of NLP, it is helpful to distinguish two functions of language, language as a live medium of communication between human beings, or between a human being and a machine, and language as the major means of storing and transmitting the world's knowledge. Roughly this is the distinction between the spoken and the written word. Much of the language research in artificial intelligence has been concerned with the former, e.g. in man-machine interactive systems, and in question-answering systems. In the latter, for example, the natural language processor acts as an interface between the human user and a structured data base. The data base has not been in question; it has for the most part been entered into the system in tabular form. The artificial intelligence task has been to interpret and act upon the natural language input produced by the user. The research described here is concerned rather with language as a storage medium, with the problem of structuring a data base which is given in natural language.

These two different foci of research activity may be illustrated by the diagram in Figure 3.1. Most artificial intelligence work in natural language processing has been located at the user end, whereas the work reported here has been located

---

*Presented in part at the Symposium on Directions in Artificial Intelligence, New York University, Dec. 6, 1974 [34], and at the Colloquium on Knowledge Delivery (Fact Retrieval) Systems Research, of the Division of Science Information of the National Science Foundation, Nov. 4, 1974.

Figure 3.1



at the data or document source end. The aim is to bring natural language data bases into such a form as to make them amenable to processing in user-directed systems for question-answering and data processing.

As an example, currently there is a great need in the medical community for computer programs that could process the information in patient records for evaluation of health care and for clinical research. This information is recorded in natural language, in dictated reports of clinic visits, in hospital discharge summaries, and the like. Although the contents of these records are limited and repetitive, their form is necessarily in natural language (except for certain parts, such as laboratory findings); e.g. no hospital administrator would dare to request physicians to substitute a multiple choice format for a dictated discharge summary. The question is thus posed as to whether there are methods for obtaining the equivalent of a predetermined format from the natural language material in cases where it is either impossible or unacceptable to impose an a priori structuring.

Another area of application is the scientific literature. In this area, again, while there are review articles every few years in very small specialized areas of scientific knowledge, no one has yet written a review article in the form of a table. Nor is it likely that that is going to happen very soon. Yet the exponential growth of the scientific and technological literature and the pressure to obtain and use available results quickly make it desirable to develop computer methods for processing the contents of the literature store, if this is at all possible. There are sufficient restrictions in the language that is used in these areas to suggest that natural language processing methods can be used to extract structures which characterize content, and in some cases in specialized areas provide the equivalent of a structured data base. The LSP group has obtained promising results in this direction working in relatively small subject matter areas, though these are relatively large compared with the areas that have so far been treated by computer programs for language analysis. Several years of work were devoted to texts in a subfield of pharmacology having to do with the cellular level action of digitalis, one of the main drugs used in cardiac therapy. The key to the initial success achieved in obtaining a structuring of natural language material is the fact that the pharmacologist in the case of the digitalis texts (and the physician in the case of medical records) is not writing in English seen as a whole language but in effect in a sublanguage of English. The limitations as to what can be sensibly said within the given field

show up as regularities in the linguistic records of the material. One can then view the problem of finding appropriate data structures as one of discovering the grammar of the sublanguage of the given scientific area.

In contrast with the grammar of a whole language we expect the grammar of a sublanguage to have categories which correlate with semantic categories in the material; that instead of obtaining such general classes as noun, or noun singular, noun plural, etc. we would obtain subclasses on a grammatical basis which would represent the objects and relations of interest in the science. In the digitalis work that will be summarized presently it turned out that this is indeed the case. Linguistic methods were used to obtain the grammar of the sublanguage, and the resulting grammatical classes were found to correspond directly to the semantic classes of the field, as recognized by a pharmacologist working in the field [13].

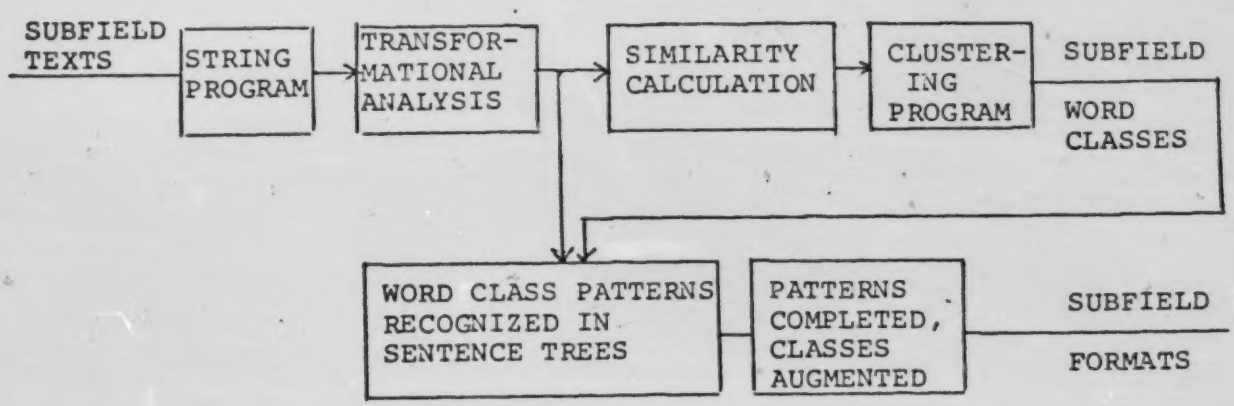
One of the reasons that this task can be viewed as a grammar-writing problem is the fact that scientists in respect to their fields display linguistic behavior; that is, certain sentences are acceptable and certain sentences are not acceptable within their field, regardless of the truth or falsity of the sentences. It is just a question of whether they are possible sentences within the field. So, for example, "the influx of sodium" and "sodium flows into the cell" are acceptable utterances in pharmacology. However, "the cell flows into sodium" is just as readily rejected by the pharmacologist as

an educated speaker of English would reject an ungrammatical sentence in the English language. It is this kind of linguistic behavior which is captured in the form of a grammar of the science sublanguage.

How is such a grammar discovered? Observe in Figure 3.2 that on the input side we have subfield texts. There are two stages of syntactic analysis, followed by two steps of statistical analysis, the result being a set of word classes that correspond to the semantic classes in the subfield. Then we have two steps of pattern recognition, the last one of which is the human stage of post-editing and putting everything together. The result is what we call a subfield information format, a kind of structuring that represents the types of information in successive sentences of the text. We will go over this step by step.

Figure 3.2

Obtaining Information Structures from Subfield Texts



Consider the first two stages of syntactic analysis. These are done by the Linguistic String Parser [Refs 23-25, 35-42, 31] which performs a segmentation or a surface analysis of the sentence followed by a second stage of transformational analysis. The type of parse obtained using string analysis greatly simplifies the following stage of transformational decomposition because the segmentation of the sentence and the labeling of strings and string-elements indicate which reverse transformations apply. In the output parse for the sentence, "This results from the slowing of the influx of potassium into the cell," shown in Fig. 3.3, the reference to "this" in line 2 appears with an indication that it is a referential to some previous item in the text; "results" has as its object another segment (in lines 4-5), whose verb is "slowing" and whose object is "the influx of potassium into the cell". Notice that each segment of this nest of embedded objects contains a verb or verb-like element with adjuncts. In line 2, "results" is a tensed verb. In line 5, "slowing" is a verb with an "ing" suffix, making it nounlike, and "influx", while it is morphologically a noun, is related to "flow into".

This hierarchy of verbs is obtained explicitly by transformational analysis. One form of transformational output (as embedded assertions under transformational operators) is shown for the example sentence in Fig. 3.4. This subpart can be readily converted to a dependency tree, as shown in Fig. 3.5. It turns out that in the parses there is already a semantic

THIS RESULTS FROM THE SLOWING OF THE INFLUX OF POTASSIUM INTO THE CELL .

PARSE 1

1; SENTENCE = INTRODUCER CENTER ENDMARK
2, .

2; ASSERTION = SA SUBJECT SA TENSE SA VERB SA OBJECT RV SA
3, ( ) RESULTS 4,

3; LN = TPOS OPOS APOS NSPOS NPOS
THIS

4; PVINGSTG = P VINGSTG
FROM 5,

5; VINGOFN = LN LVINGR #OF# OBJECT RV SA
6, SLOWING OF 7, INFLUX 8,

6; LN = TPOS OPOS APOS NSPOS NPOS
THE

7; LN = TPOS OPOS APOS NSPOS NPOS
THE

8; PN = LP P NSTGO
OF POTASSIUM 9,

9; PN = LP P NSTGO
INTO 10, CELL

10, LN = TPOS OPOS APOS NSPOS NPOS
THE

Figure 3.4

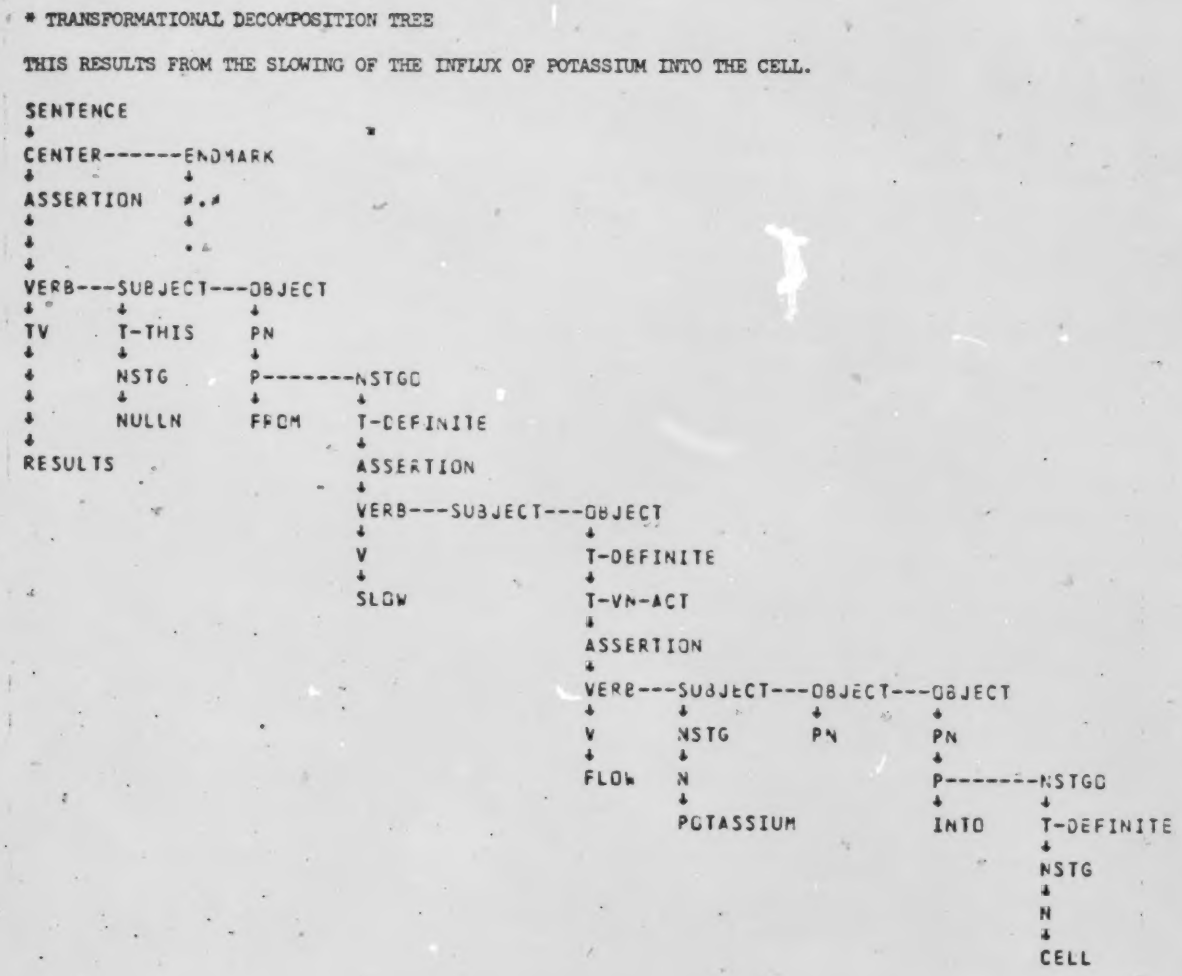
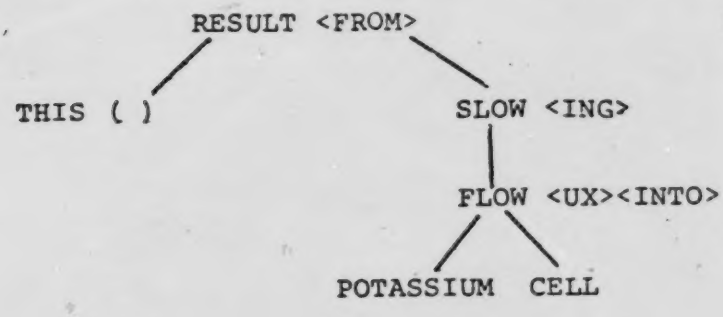

* TRANSFORMATIONAL DECOMPOSITION TREE
THIS RESULTS FROM THE SLOWING OF THE INFLUX OF POTASSIUM INTO THE CELL.
SENTENCE
CENTER
ENDMARK
ASSERTION
#.#
.
VERB
SUBJECT
OBJECT
TV
T-THIS
PN
NSTG
P
NSTGO
NULLN
FROM
T-DEFINITE
RESULTS
ASSERTION
VERB
SUBJECT
OBJECT
V
T-DEFINITE
SLOW
T-VN-ACT
ASSERTION
VERB
SUBJECT
OBJECT
OBJECT
V
NSTG
PN
PN
FLOW
N
P
NSTGO
POTASSIUM
INTO
T-DEFINITE
NSTG
N
CELL

Figure 3.5

Transformational Dependency Tree



separation of what one might call the object language of the science from the meta-language. The lower parts of the parse tree contain the events and the objects of interest in the field, and the upper parts of the parse tree the human relations to these events, in such expressions as "it was assumed that", or "we observed that". Within the lower range there is a hierarchy of verbs. The lowest level here are verbs of qualitative events; then come certain quantitative verbs and then certain causal or sentence connecting verbs. This verb hierarchy is made explicit by the stage of transformational analysis illustrated by Figs. 3.4 and 3.5. In a transformational decomposition tree of this type each verb is seen to have certain arguments, e.g. "flow into" with its arguments "potassium" (subject) and "cell" (object). This is operated upon by "slow", operated on in turn by "results from" which also has another argument of a sentential type.

This type of output from the first two stages of syntactic analysis leaves us a tree on which we can do statistical clustering operations to obtain the noun and verb classes. The

trick here is to cluster words that are similar in their grammatical position within the tree, i.e. have the same word as their operator or argument. For example, we read off from the tree in Figure 3.5 (and from several sentence trees not shown) certain word triples, as shown in Figure 3.6. At the top of Figure 3.6 there are two instances of "move" and one of "flow", all of them having "potassium" as their first argument, or subject. "Flow" and "move" are similar in several respects here. First of all, they have the same word as their first argument; that is, they have a similarity in the transformational trees by having the same word in the same argument position. Nouns also, e.g. "potassium" and "sodium", are found to be similar by the same type of criterion. That is, they have the same verb as their operator while they are in a given argument position.

Figure 3.6

Similarity Coefficients

FLOW <UX><INTO>	POTASSIUM	CELL
MOVE <MENT><ACROSS>	POTASSIUM	MEMBRANE
MOVE <MENT><INTO>	POTASSIUM	CELL

MOVE similar to FLOW

FLOW <UX><INTO>	POTASSIUM	CELL
FLOW <UX><OUT OF>	SODIUM	CELL

POTASSIUM similar to SODIUM

We are now in a position to compare quantitatively the similarity of occurrences of words. We have seen qualitatively how words in particular grammatical classes are similar in a

tree; now we will do it quantitatively. We will compare every pair of words in the entire corpus with respect to six possible grammatical relations: namely the first, second, or third argument of the other word if that word is an operator, or the operator where the other word is the first, second, or third argument. For example, in Figure 3.7, we see the data for "flow" and "move", taken from Figure 3.6. "Move" had two occurrences with "potassium" as subject. "Flow" has one such occurrence. The calculation is made on the very large (sparse) matrix of grammatically differentiated word pairings. Each word is assigned a characteristic (normalized) vector based on tables of the type illustrated in Figure 3.7, showing the number of "similar" occurrences the word has vis-à-vis all other words. We obtain a similarity coefficient, that is, a numerical measure of

Figure 3.7

Computing Word Similarities

WORD PAIR FREQUENCIES

	POTASSIUM AS SUBJECT	POTASSIUM AS OBJECT	CELL AS OBJECT	
FLOW	1	0	1	ALL WORD PAIRS
MOVE	2	0	1	EACH PAIR IN 6 POSSIBLE GRAMMATICAL RELATIONS

NORMALIZED FLOW VECTOR $\vec{F}$

NORMALIZED MOVE VECTOR $\vec{M}$ (6n DIMENSIONS)

SIMILARITY COEFFICIENT = $\vec{F} \cdot \vec{M}$

how similar each two words are in our text corpus, by doing a vector multiplication.

We thus obtain a number that represents the similarity of any two words in the corpus, as illustrated in Figure 3.8. The slashed squares contain similarity coefficients that are above threshold. The threshold here is .27. Every i,j square contains the calculated similarity between the two words in the ith row and jth column. We expect words that have a similarity coefficient greater than the threshold to form a cluster by a clustering algorithm, as illustrated in Figure 3.9.

Figure 3.8

Similarity Coefficients

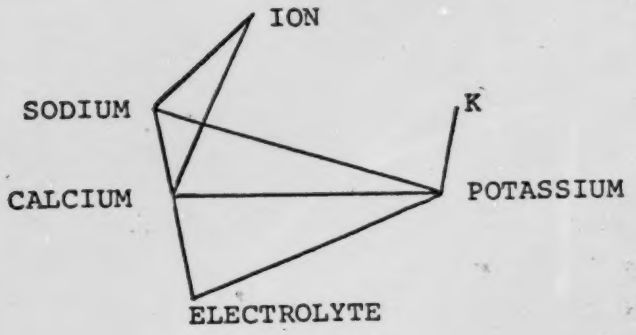
	K	POTASSIUM	SODIUM	ELECTROLYTE	ION	CALCIUM
K		.457	.151	.075	.110	.220
POTASSIUM			.365	.348	.147	.514
SODIUM				.148	.449	.488
ELECTROLYTE					.072	.312
ION						.286
CALCIUM						

Threshold = .27

In Figure 3.9 we have replaced the slashed squares of Figure 3.8 by lines, showing that the words that have a similarity coefficient higher than the threshold form a cluster. And as the figure illustrates, the cluster of Figure 3.9 groups

Figure 3.9

Words with High Similarity Coefficient Form a Cluster



together words which have a certain semantic property in common; the cluster contains ion names plus classifications of ion names. They form a cluster because of their similar type of occurrence in the textual material. The clustering algorithm operates by adding one word at a time to a started cluster. A word is added to a cluster if the average of its similarity coefficient to all of the others in the cluster is above threshold. (Details are given in [32]).

We may now turn to the results obtained by the clustering program operating on the digitalis material. This experiment involved about 400 sentences from a much larger corpus, taken from four or five texts, not selected in any particular way, except that they were part of the set of digitalis texts that we had analyzed manually in an earlier study. A look at the clustering output in detail (Figs. 3.10, 3.11) demonstrates that the clusters are semantically coherent, so much so that we were able easily to name them. It should be mentioned that the

Figure 3.10

CLUSTERING OUTPUT

Run 11.13.74
T = 0.27

NOUN CLASSES

CG CLASS

CG
DIGITALIS
OUABAIN
DRUG
STROPHANTHIDIN
STROPHANTHIDIN 3 BROMOACETATE
STROPHANTHIN
CARDIOTONIC GLYCOSIDE
COMPOUND
INHIBITOR
ERYTHROPHLEUM ALKALOID
INHIBITOR
AGENT

CATION CLASS

CALCIUM
CA^{++}
CA
POTASSIUM
K
SODIUM
NA^{+}
ION
ELECTROLYTE
GLUCOSE
ION
K^{+}

ENZYME CLASS

$NA^{+}K^{+}$ ATPase
ENZYME
ATPase

PROTEIN CLASS

FIBER PROTEIN
PROTEIN ACTOMYOSIN
CARDIAC

SR CLASS

SR
SARCOPLASMIC RETICULUM

MUSCLE CLASS

MUSCLE
HEART MUSCLE
MUSCLE
VENTRICLE

CELL-TISSUE CLASS

CELL
MYOCARDIUM

ADP CLASS

ADP
E1

SPACE CLASS

SPACE
MILIEU

Figure 3.11

CLUSTERING OUTPUT

VERB CLASSES

RUN 11.13.74
T = 0.27

V_{IC} (MOVE)

MOVE
TURNOVER
INTRA
EXTRA
CONCENTRATE
FLOW

V_{CI} (CONTAIN)

CONTAIN
LOSE

V_{M} (EXCITE)

EXCITE
DEPOLARIZE

V_{Q} (INCREASE)

INCREASE
DECREASE
CHANGE
AUGMENT
INCREASE

V_{SS} (CAUSE)

CAUSE
INDUCE
PRODUCE
LINK
INTERFERE
ACT
AFFECT
RELATE
TOXIC

INFLUENCE
INHIBIT
STIMULATE
AFFECT
CONCENTRATE
ACT
REVERSE
INDUCE
TOXIC
PENETRATE

DEMONSTRATE
SIMILAR
CAUSE
RELATE
DUE TO
PRODUCE
LINK

OPPOSE
DIVERGE
SIMILAR

EFFECT
PRODUCE

DISSOCIATE
RELATE

V_{S} (UNNAMED)

REPORT
OBSERVE

TAKE
TREAT

TRANSPORT
EXCHANGE

AUGMENT
IMPROVE

REDUCE
INFLUENCE

DECREASE
MEASURE

clustering program generates overlapping clusters which are then merged by a simple algorithm that merges those overlapping clusters that have a 2/3 intersection. Figures 3.10 and 3.11 show the output from the merging algorithm, i.e., the merged clusters from a particular run. CG stands for "cardiac glycoside"; the words that fell into the merged cluster labeled CG are with one exception cardiac glycosides or drug classifiers. There is one interesting exception; the erythrophleum alkaloids are not glycosides, but they are a set of related compounds which in a particular text that we were analyzing happened to be compared with the glycosides and therefore they fell into the CG cluster. The output shows that there is a certain semantic coherence to the class generated by the clustering program. It does not mean that the machine is perfect and will make a perfect semantic class, but it shows that the machine does not put unrelated words into one class, because it is based on distributional similarities of words. It should be noted that the word classes in Figures 3.10 and 3.11 are not selected from the data. They show the complete computer output for the run in question.

The verb classes shown in Figure 3.11 are interesting. Since the nouns have been clustered with respect to the verbs and the verbs with respect to the nouns we have labeled some of the verb classes in terms of their relational status to nouns. For example, V_{IC} are the verbs that related ion words to cell words. Notice that they are verbs of motion, or of arrested motion. "Moves", "turnover", "concentrate", "flow". ("Intra",

"extra" are treated as verbs.) The V_{CI} are just the reverse; they relate cell words to ion words, e.g. "contain" as in "the cell contains sodium". Then we have a level of quantitative verbs: "increase," "decrease", "change", "augment", "increase". The parallel columns indicate that the clusters were not merged by the merging algorithm. Since the merging algorithm requires a 2/3 overlap, there was no merging of pairs; the last step of putting these clusters together was done by simple alignment. Notice that the computer has distinguished six different kinds of sentence-connecting clusters. We don't know as yet what this really represents. When we did a manual analysis of this material we put all of what seemed to be causal verbs, i.e. the relational verbs between events, into a single class, e.g. "cause", "induce", "produce", "link", "interfere", "act" ("act" is "act on", and "toxic" is "toxify" in this case), "influence", "inhibit", "stimulate", and so forth. All of these represent relations among events in the physiological world, in particular, among events that are initiated by the drug actions. And as can be seen there are nevertheless differences recognized by the computer program, which would be very interesting to explore; the clustering algorithm saw the members of these subgroups as being more closely associated with each other than with the words in the other subgroups.

To be complete, we have to check what was missed by the algorithm as well as what was obtained. We did this by analyzing the same material manually, obtaining the semantic classes and

and formulas by observation and by doing a rough distributional analysis. We checked the validity of the manually obtained classes with a pharmacologist. Then we compared them with the computer output. It turned out that in each of the given categories the computer classes contained all of the high frequency occurring elements that were in the manual classes. For example, in Figure 3.12, the manually obtained cardiac glycoside class contains the same words that were obtained by the computer, plus some others, which have frequencies of occurrence that taper down toward singular occurrence. The computer was able to collect all of the similar words which had a high frequency and missed only those which were of low frequency. In an application, this situation could be corrected by a larger corpus or augmenting the computer-generated class (treated as a nucleus) manually.

A similar result is seen in Figure 3.13 with the cation class. The words in the computer class accounted for 96 percent of the cation occurrences in operator-argument pairs, but missed a few low frequency words which would have fitted naturally into the class.

At this point we may refer back to the diagram in Figure 3.2 for orientation. We have discussed so far the two steps of syntactic analysis and the two steps of statistical analysis which together provide subfield semantic word classes. Now we want to go back into the sentence and obtain with the aid of these classes the patterns of word class co-occurrence which can lead

Figure 3.12

CG CLASS

COMPUTER	MANUAL	NO. OF PAIR OCCURRENCES	
CG	CG	156	
DIGITALIS	DIGITALIS	118	
OUABAIN	OUABAIN	70	
DRUG	DRUG	15	
STROPHANTHIDIN	STROPHANTHIDIN	5	89%
STROPHANTHIN	STROPHANTHIN	4	
STROPHANTHIDIN 3 BROMOACETATE	STROPHANTHIDIN 3 BROMOACETATE	4	
CARDIOTONIC GLYCOSIDE	CARDIOTONIC GLYCOSIDE	3	
COMPOUND	COMPOUND	7	
INHIBITOR	INHIBITOR	5	
	GLYCOSIDE	11	
	AGENT	8	
	DIGOXIN	7	
	ACETYL STROPHANTHIDIN	7	
	CARDIOACTIVE GLYCOSIDE	6	
	DIGITALIS GLYCOSIDE	6	
	DIGITOXIGENIN	3	
	STROPHANTHOSIDE	2	
	CARDIAC GLYCOSIDE	2	
	DIGITOXIN	1	
	STROPHANTHIN K	1	
	DIGITALIS COMPOUND	1	
ERYTHROPHLEUM ALKALOID 6			
		442	

RUN 11.13.74
T = 0.27

Figure 3.13

CATION CLASS

COMPUTER	MANUAL	NO. OF PAIR OCCURRENCES	
CALCIUM	CALCIUM	101	
POTASSIUM	POTASSIUM	90	
SODIUM	SODIUM	53	
CA^{++}	CA^{++}	48	
CA	CA	30	96%
K	K	29	
ELECTROLYTE	ELECTROLYTE	17	
ION	ION	15	
NA^{+}	NA^{+}	11	
GLUCOSE		7	
	K^{+}	6	
	CATION	3	
	MG^{++}	3	
	MAGNESIUM	3	
	NA	3	
		412	

RUN 11.13.74
T = 0.27

us to information structures for the material. It should be clear from the way the trees were built and the way the words were clustered from the trees that once the major subfield word classes are obtained, we can plug them back into the sentence in the positions from which they were clustered so as to obtain patterns of word class co-occurrence. By so doing, we will get either fragmentary or complete patterns in terms of word classes, in all sentences containing words which appear in clusters. (See also the discussion of Fig. 2.5, above.) At this point some human intellectual work is necessary to put together fragments into more complete structures. We obtain structures which seem to account quite regularly for units of content which appear frequently in part or in whole (and in many different grammatical forms) in many text sentences. They constitute alignments of like information and as such have been called "information formats" [13, 14].

As an example, Fig. 3.14 shows the information format for the initial sentence shown in Figure 3.3. The word "this" takes up a whole format line. It is connected to the next format line by a conjunction which here is the sentence connecting verb, "results" with its preposition "from". "Results from" requires two nominalized sentences or their stand-ins as its arguments. This fact determines that "this" in line 5.1 stands for an event and justifies its taking up a format line. The second format line (labeled 5.2) in Fig. 3.14 contains two levels of embedding: the innermost part represents the elementary

Figure 3.14

Format of S_5

	V_Q	S_E: N_1	S_E: V	S_E: P N_2	D_S	CONJ
5.1	←	{4,} THIS			→	RESULTS FROM
5.2	SLOWS	POTASSIUM	INFLUX	INTO THE CELL		

SENTENCE: THIS RESULTS FROM THE SLOWING OF THE INFLUX OF POTASSIUM INTO THE CELL.

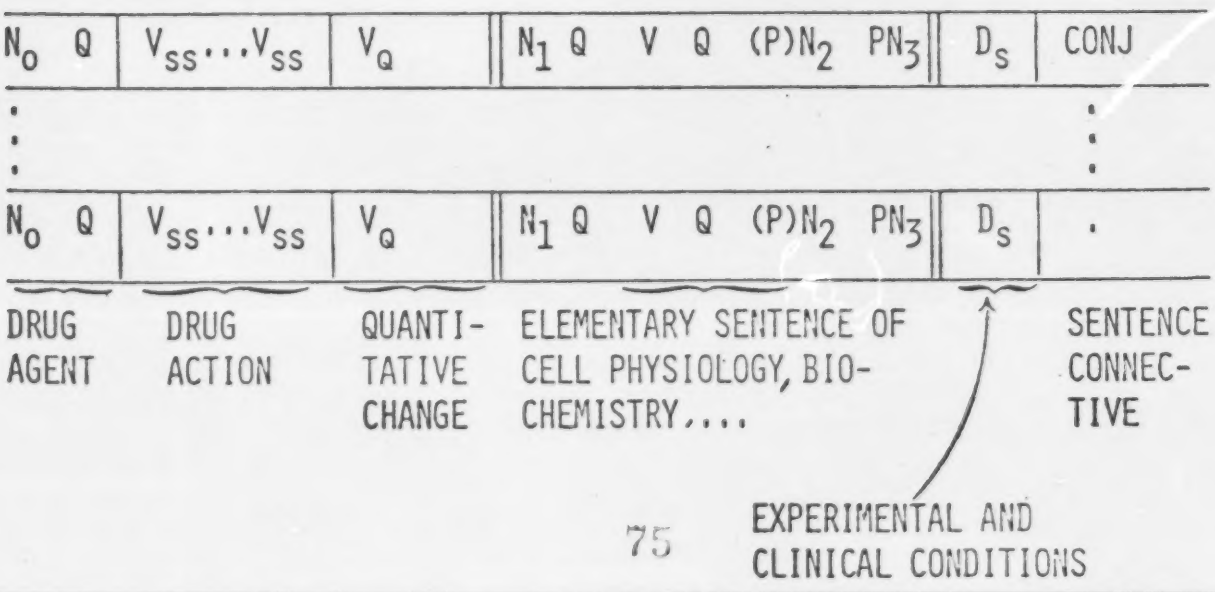
sentence: "Potassium flows into the cell," operated upon by the verb "slows". This is really just the same sentence that was seen before, but it is a case now of a type of pattern which we recognize in very many sentences of the text. The format for this rather complicated field of the cellular action of digitalis has a very repetitive structure when seen in terms of the types of information contained in factual portions of text sentences.

Information formats do not characterize the course of an argument or other features of the discourse structure within a given text. The procedures described here can be considered successful if they enable us to establish a typology for the kinds of information, i.e. the major classes of interest and the types of relations that are of interest, within a field. The results may also be useful in further studies of text structure.

In the case of the digitalis subfield an overall repeating pattern, of which there were many subtypes, was found in fact-reporting portions of texts (Fig. 3.15, repeating Fig. 2.6). The repeating pattern is one where there is some elementary qualitative sentence of either cell physiology or biochemistry at the deepest embedded level, a sentence which has a concrete verb, and concrete nouns as its subject and object. These elementary sentences are operated upon optionally by some quantitative material; and this in turn has some sequence of causal verbs operating on it, appearing one after the other with an initiating drug agent as the final subject. This format was checked by going through a number of articles and seeing if they could be formatted according to this pattern, and indeed the large majority of the sentences had this kind of pattern, with different detailed sentences within it, but all of them cases of the word classes (and some individual words) in particular relations.

Figure 3.15

Partial Text Format for Pharmacological Subfield



A few last words about the implementation of the components in the above procedures. The linguistic string parser has been operative in one programmed version or another since 1966 [see references cited above]. Here we might note those features which make possible its application to semantic problems in text analysis, First, it has a very large and detailed grammar of English, absolutely necessary for any real text analysis. Secondly it has an extensive treatment of coordinate conjunctions and comparatives, also without which you cannot do very much real text analysis. Thirdly it now utilizes a programming language which has been developed at the NYU Linguistic String Project for writing the necessary grammatical and semantic restrictions, and which also makes the stating of reverse transformations quite straightforward. The implementation of the current parser and of the programming language and the transformational mechanism have been done by Ralph Grishman. The transformations are being slowly entered into the system, chiefly by Jerry Hobbs and Carol Raze. The treatment of coordinate and comparative constructions, both parsing and transformations, is the work of Carol Raze. The similarity calculation and clustering programs were written by Ralph Grishman in cooperation with Lynette Hirschman, who also did the linguistic analysis.

It should be noted that in the clustering experiment just described the transformational component of the computer grammar was not yet rich enough to handle the text sentences,

so the trees for that stage were manually done. The trees were drawn by hand for the input sentences, but used only transformations that were being added to the grammar.

Finally, how can these results be used? What has been described thusfar is a discovery procedure for the word classes and data structures of a subfield. Figure 3.16 combines a review of the procedures with indications as to how some of the components could be used in applications. The components shown in the top part of the diagram obtain the appropriate word classes and data structures; those shown in the bottom line process new texts. Products of component procedures and some of their potential uses are indicated.

Not shown in Fig. 3.16 is the use of subfield word classes to resolve syntactic ambiguity. Some of the problems that are difficult in parsing can be solved with the aid of a sublanguage grammar. The subfield word classes constitute a semantic tool for resolving ambiguities that arise in parsing subfield texts, as illustrated in Figure 3.17. In interpreting the sentence in Figure 3.17, it is clear to the reader that changes in the ionic milieu of cells is what is produced by digitalis, rather than that the cells are produced by digitalis. But this ambiguity cannot be resolved without semantic constraints that are quite specific to the subfield (and are not universal semantic constraints in the language). In this case, in the digitalis sublanguage the verb "produced" never occurs in the sublanguage with an anatomical word as its second argument,

Fig. 3.16

Subfield Semantic Structures: Discovery and Use

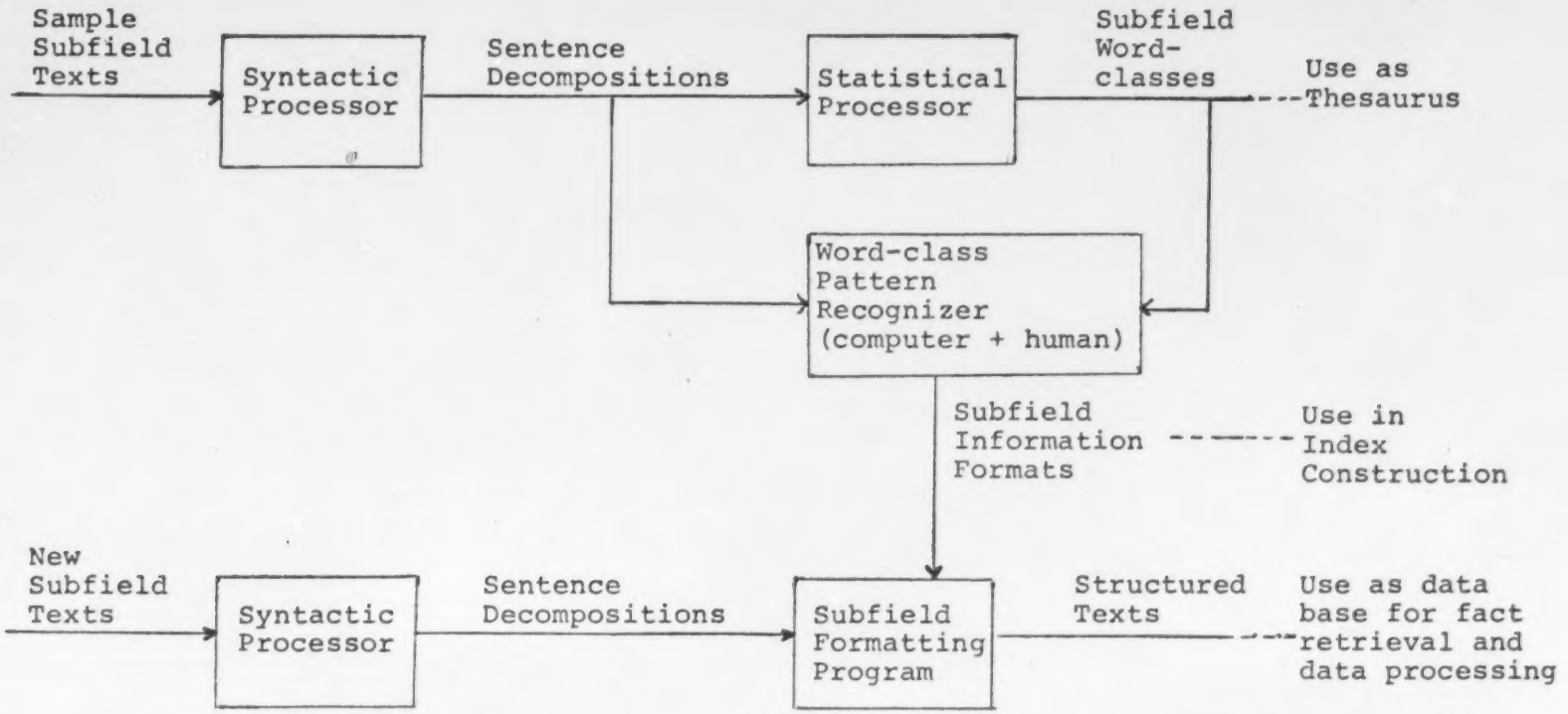
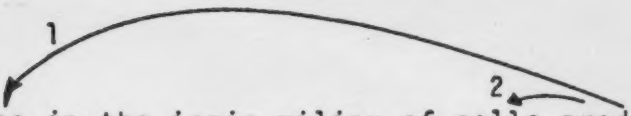


Figure 3.17

SYNTACTIC AMBIGUITIES RESOLVED BY

SUBLANGUAGE CONSTRAINTS



Changes in the ionic milieu of cells produced by digitalis have been known for many years.

1.	Digitalis produces changes	YES
2.	Digitalis produces cells	NO

i.e. there is no "produce cells", "produce membranes". The second interpretation is far less likely than the first because it does not fit any of the co-occurrence patterns that have been established for the sublanguage.

4. APPLICATIONS

Computerized language processing at its most successful will not replace existing methods of information retrieval, but may supply tools for improving present services and, perhaps more significantly, open the way to new services which are only conceivable in the wake of advanced linguistic techniques.

4.1 GENERATING WORD LISTS

Under the rubric of tools would be new approaches to thesaurus generation based on linguistic analysis of texts. Experimental results obtained in generating word sets from transformationally analyzed pharmacology sentences are quite encouraging in this respect. The experiment described in section 3 demonstrated that a statistical treatment of linguistic tokens could produce semantically sharp and informationally relevant word classes when the clustering took into account the grammatical relations of words in the sentences. To summarize, the sentences in the experiment were all from texts treating the mechanisms of action of digitalis; "Methods" sections were excluded but the sentences were not otherwise especially selected. Two words were considered similar if they occurred in the same argument position under the same operator (e.g. as subject of a certain verb) or as operator on the same operand (e.g. as verb with the same noun object) in the transformational decomposition of a sentence. Similarity coefficients SC were computed for all possible pairs

of words, and clusters were built up one word at a time; a word was added to a cluster if the average of its SC with each other word in the cluster exceeded a threshold value. Clusters with 2/3 overlap were merged to produce the final word classes. These classes turned out to be semantically coherent, and the set of word classes as a whole included the main types recognized by a cardiologist for the same material.

To check the coverage obtained by the computerized procedure, all the words in the corpus were manually classified and the resulting classes were compared. Two tables showing the overlap of the computer and manual classes were given in Figs. 3.12, 3.13. The former is repeated with more detail in Fig. 4.1, which shows the computer class of drug words (cardiac glycosides CG) compared with the manually obtained class. The computer class contains all the high frequency words which appeared in the manual class and only one other word. The low frequency terms which were missed might be captured in a larger corpus. (It was surprising how much was obtained with such a small set of sentences, ca. 400.) Or, once the important classes were established, they could be supplemented by manually examining word lists or consulting an expert.

The algorithm clearly works well in particular subject matter areas where not only do the texts have overlapping vocabularies, but usage has become routinized. Whether similar results can be obtained over broader subject matter areas remains to be investigated. It should also be noted that the

Figure 4.1

CG CLASS COMPUTER	MANUAL	No. OCCURRENCES IN PAIRS†	
CG	CG	156	
digitalis	dititalis	118	
ouabain	ouabain	70	
drug	drug	15	
agent	agent	8	$\frac{395}{442} = 89\%$
strophanthidin	strophanthidin	5	
strophanthidin 3 bromoacetate	strophanthidin 3 bromoacetate	4	
strophanthin	strophanthin	4	
cardiotonic glycoside	cardiotonic glycoside	3	
compound	compound	7	
inhibitor	inhibitor	5	
*erythrophleum alkaloid		*6	
	glycoside	11	
	digoxin	7	
	acetyl strophanthidin	7	
	cardioactive glycoside	6	
	digitalis glycoside	6	
	digitoxigenin	3	
	sprophanthoside	2	
	cardiac glycoside	2	
	digitoxin	1	
	digitalis compound	1	
	strophanthin K	1	
		442	

**Erythrophleum alkaloid* does not belong in the CG class; it is a drug whose effect is compared to that of the cardiac glycosides.

Agent, *drug* and *compound* are classifiers for words of the CG class, as well as of the more general DRUG class. *Inhibitor* is also a classifier, which classifies according to function.

†An occurrence of a word either as the operator or operand in a pair. Pair-occurrences are more numerous than text occurrences for several reasons. Recoverably zeroed material is reconstructed and contributes to pair formation. Also each operator can appear in a pair as the operand of its operator, as well as with each one of its arguments. (Thus a two-argument verb can appear in three pairs.) For concrete nouns however this does not occur, and the pair-occurrences correlate more closely with the number of actual occurrences in the text.

input trees in this experiment were not obtained by computerized decomposition of sentences, though only standard English transformations which were in the process of being programmed were used. The use of such clustering algorithms on an extensive scale would of course depend on obtaining the input from computer sources, and one should not minimize the problems of real text processing. The very complexity of most science sentences even if they are not ambiguous is forbidding. In addition, the problem of ambiguity is real. In applying the clustering program to computer outputs of text decompositions, it would be necessary to select an initial set of unambiguous outputs (selected by hand or by skipping over sentences with multiple outputs) and to use them in a bootstrap operation to obtain others. Presumably, the feeding back of selectional constraints based on results from the initial set of outputs would aid in disambiguating a number of the remaining outputs. However, this step may not be crucial. The amount of repetition of significant operator-argument pairs is so great that even if only a small percentage of the text sentences can be used for clustering, the results are still likely to be creditable.

4.2 GENERATING TEXT DESCRIPTORS

In the class of tools, or adjuncts to retrieval, is the interesting possibility raised by Sparck Jones and Kay concerning the automatic generation of structured index descriptions. In the concluding paragraph of the chapter on Syntax in Information Retrieval [5:p. 119] they write:

> it is intriguing to imagine a system in which document descriptions using links and roles would be obtained automatically from deep structures produced by an advanced transformational analysis technique. This would be a difficult enterprise, but there is no doubt that it would blend theories and methods of interest in both linguistics and information science and the results could hardly fail to be instructive.

To put this hypothesis to an immediate informal test, a small experiment was run on the pharmacology corpus for which we had both the clustered word classes and the linearized transformational trees. All triples consisting of an operator and its arguments were generated. Every word which was a member of one of the three main noun classes DRUG, ENZYME, ION* was replaced by the name of the class to which it belonged. Triple occurrences were then counted and all those with a frequency of 5 or more were printed. The printed list was then culled for all triples whose arguments were nouns. (Note: nominalized verbs like uptake, etc. appear as verbs, not nouns, in the decompositions.) The results are shown in Fig. 4.2.

It is, of course, hard to judge the output of Fig. 4.2 as a representation for a set of texts which are not available to the reader. It is the case that these articles (and others in the field) were primarily concerned at this time (mainly late 60's) with the possible effects of digitalis on ion transport and on the activity of the enzyme ATPase (in particular

---

*Pharmacology Noun classes for run of 3/10/76.
DRUG CLASS: STROPHANTHIDIN 3 BRO, DRUG, DIGITALIS, STROPHANTHIDIN, OUABAIN, AGENT, GLYCOSIDE, CG, ERYTHROPHLEUM ALKALO, QUINIDINE, CHLORPROMAZINE, STROPHANTHIN, INHIBITOR, COMPOUND
ENZYME CLASS: NA+ K+ ATPASE, ENZYME, ATPASE
ION CLASS: K+, NA+, ION, SODIUM, CALCIUM, POTASSIUM, GLUCOSE, CA++, K, ELECTROLYTE, NA, CA.

Figure 4.2
Pharmacology Operator-Argument Triples‡

ARG1	OP	ARG2	f/f'*
DRUG	affect	ENZYME	26/276
	inhibit		16/95
	stimulate		7/52
	bind		6/46
			55/469
ION	activate	ENZYME	8/17
ION	concentrate	cell	15/121
	intra		16/20
	level (has)		6/12
	move		6/20
	concentrate	tissue	6/121
	flow	muscle	7/51
			56/114
cell	contain	ION	19/50
myocardium	contain		10/50
	lose		5/28
SR	find		9/46
	take (up)		23/41
			66/165
actin	interact	myosin	7/14
fiber	extract	glycerol	6/9
fiber	contract		6/105
muscle			6/105
protein			9/105
			21/105

‡shown here for noun arguments only and for triples occurring with frequency $f \geq 5$. Capitalized words stand for noun classes. The table includes pairs due to 1-argument verbs (e.g. contract).

*f' = total no. of occurrences of operator in the corpus.

Na^{+}-K^{+}-activated ATPase) as possible explanations for the ultimate effect of digitalis in enhancing the contractility of heart muscle. There is much discussion in the texts of ion movements in the cell, and the beginning of the discussion (which becomes more prominent in later papers) of events at the level of the contractile proteins, in particular those involving Ca^{++} movements in relation to actin and myosin. The triples are, then, a fair quick summary of what the articles deal with (though, of course, not of what they say).

Some additional comments are in order. Note that the f/f' ratios of the first group are much lower than those in the other groups. This is because the verbs in the first group are mainly sentence-connecting verbs rather than noun-connecting verbs, i.e. they carry the posited connection between events. These connections are not caught by triples because a verbal argument without its own arguments gives little information. For example, the computer output contained the group of triples:

			f/f'
	inhibit	act	5/95
DRUG	affect	take (up)	12/276
	produce	increase	5/44

which is really the first part of a larger n-tuple linking this group with some of the triples in other groups, e.g. combining a line of the above group with one from the fourth group in Fig. 4.2 would give the quintuple:

DRUG affect SR take(up) ION

This corresponds to sentences such as "Digitalis may have an indirect effect on the Ca^{++} uptake of skeletal SR," or sentence

parts like "the effect of the cardiac glycosides on the uptake of calcium by the sarcoplasmic reticulum." We see that this quintuple begins to resemble one of the information formats found in this material (Fig. 2.5, 2.6). There is even hope that with a large enough corpus the computer programs could discover such patterns in the textual material.

It appears, then, that triples consisting of an operator and its noun argument(s), obtained from transformational decomposition trees of text sentences, are a promising source of structured index terms. In addition, it may even be possible to recognize larger patterns as a step toward developing data structures for use in fact retrieval and in processing the textual content.

4.3 STRUCTURING NATURAL LANGUAGE DATA BASES

What we cannot do quite yet in the case of the journal literature, namely transform the natural language form of the data into structures suitable for computerized information processing, appears to be more immediately feasible in natural language data bases of a more restricted kind, such as are found in reports and records, e.g. weather reports, medical records of various types, program specifications, special purpose files in business and industry as well as in science. Several factors conspire to make this one of the most promising areas currently for applying NLP techniques, chiefly:

(a) From the NLP side, the language problems are easier to handle: In such material, the sentences (or sentence fragments)

are on the whole shorter than in full texts; sentence connectives are fewer and simpler; pronominal reference is limited; vocabulary is limited, and most important, it is used in a regular unambiguous way.

(b) From the user side, the need for such natural language processing is large. Files and records are growing, and while many of them can be assigned an overall format, the entries under the structured headings are for the most part unavoidably in natural language. At the same time many offices, hospitals, services, etc. are computerizing their files for quick input-output or other transactional purposes, so that the natural language data store is available for processing, if the programs for such processing can be devised.

The combination of the developing need, the developing natural language data bases, and the developing language techniques is an auspicious setting for a cooperative effort at developing new services. Sparck Jones and Kay suggest in their conclusion [5:p. 198] that in the task of relating linguistics and documentation more effectively, "the onus is on the documentalist to provide a proper specification of what he needs to satisfy his retrieval objectives." While clear specifications are certainly needed, innovations employing NLP techniques require that the computational linguist first demonstrate what can be done, and work with the documentalist (or information specialist in a given area) to specify new objectives. The area of natural language data bases is a good place to start since the user and

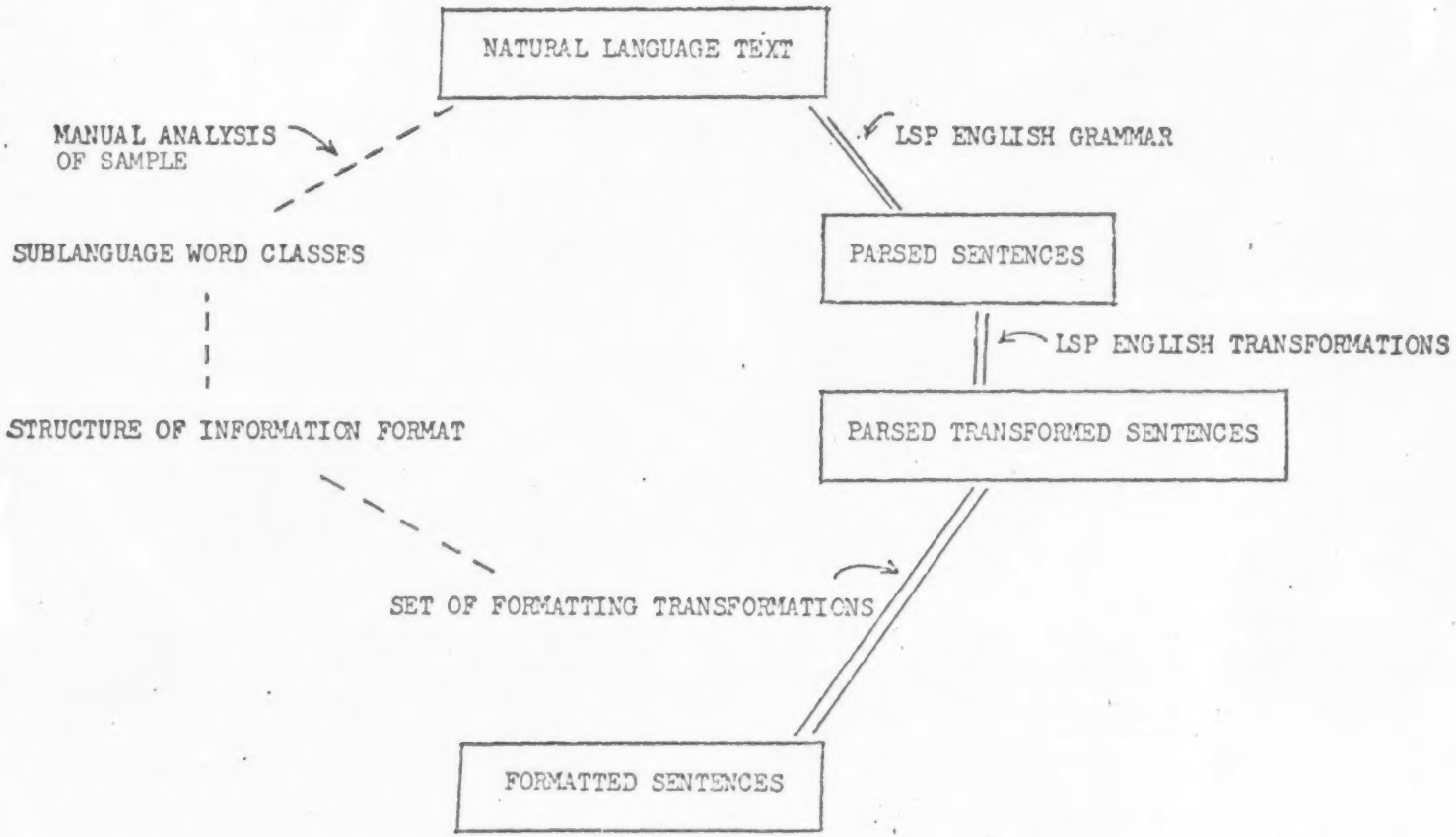
data specialist are close at hand--sometimes even the same individual, and cooperation is thus easier to achieve.

Work on the computerized structuring of information in natural language records is just beginning. Some work on weather reports is being done in Canada in the framework of mechanical translation [43]. The example provided below is from work on medical records, in particular a small corpus of X-ray reports on patients who were being followed after cancer surgery. The material ranged from simple one or two word entries like "none," "not done" to textlets like the following:

> CHEST X-RAY ON 10-14-69 SHOWED MARKED REACTIONS SUPERIORLY OF THE LEFT HILUM WITH CONSIDERABLE PLEURAL REACTION AND/OR LOCULATED FLUID IN THIS AREA IS SEEN . THIS, HOWEVER, HAS REMAINED UNCHANGED OVER THE PAST FOUR MONTHS . THE HEART REMAINS NORMAL IN SIZE AND CONFIGURATION . NO EVIDENCE FOR PULMONARY OR BONY METASTASIS .

The experiment proceeded according to the flow chart in Fig. 4.3. The labeled broken lines represent manual steps to obtain the appropriate word classes and formats (the clustering program was not used in this experiment), and the double dark lines represent steps in computer processing, equipped as indicated with a grammar and transformations. A rather detailed information format was developed, shown with one instance of a formatted entry in Fig. 4.4. The computer parse and computer format for this entry are appended in Figs. 4.5a, b and 4.6.

Figure 4.3


NATURAL LANGUAGE TEXT
MANUAL ANALYSIS OF SAMPLE
LSP ENGLISH GRAMMAR
SUBLANGUAGE WORD CLASSES
PARSED SENTENCES
LSP ENGLISH TRANSFORMATIONS
STRUCTURE OF INFORMATION FORMAT
PARSED TRANSFORMED SENTENCES
SET OF FORMATTING TRANSFORMATIONS
FORMATTED SENTENCES

FORMAT OF BROSS 1.1.2

BROSS 1.1.2 3-28-68 flat plate of abdomen--mild degenerative changes and osteoporosis of the sacral spine.

TEST					FINDING								CONN
NO-TEST	TESTN	TESTLOC	VERB-DONE	TESTDATE	NEG	VERB-ELEMENT		CHANGE	STATUS	MED-FIND.	REGION			
						BE-SHOW	INDIC.				POS	PT-BODY	STRUC.	
	(FLAT) PLATE	(OF THE) ABDOMEN		3-28-68		--		(MILD DEGENERATIVE) CHANGES						AND
	(FLAT) PLATE	(OF THE) ABDOMEN		3-28-68		--				OSTEOPOROSIS	OF	(THE SACRAL) SPINE		

Adjuncts are written in ( ); left adjuncts above the entry, right adjuncts below the entry.

Figure 4.4

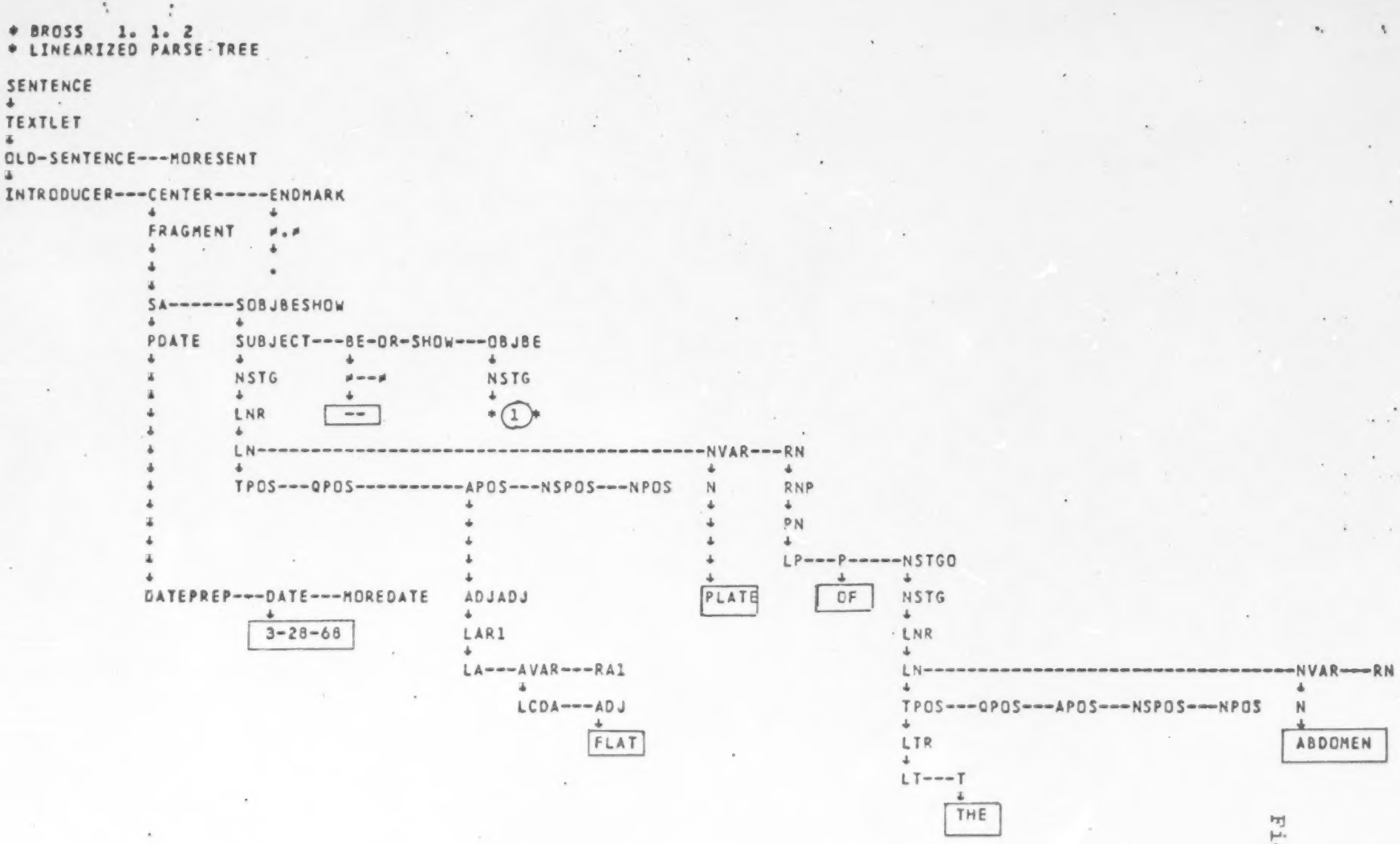

* BROSS 1. 1. 2
* LINEARIZED PARSE-TREE
SENTENCE
TEXTLET
OLD-SENTENCE---MORESENT
INTRODUCER---CENTER-----ENDMARK
FRAGMENT .
SA------SOBJBESHOW
PDATE SUBJECT---BE-OR-SHOW---OBJBE
NSTG #--# NSTG
LNR -- *(1)*
LN---NVAR---RN
TPOS---QPOS----------APOS---NSPOS---NPOS N RNP
PN
LP---P-----NSTGO
DATEPREP---DATE---MOREDATE ADJADJ PLATE OF NSTG
3-28-68 LAR1
LNR
LA---AVAR---RA1 LN---NVAR---RN
LCDA---ADJ TPOS---QPOS---APOS---NSPOS---NPOS N
FLAT LTR ABDOMEN
LT---T
THE

Figure 4.5a

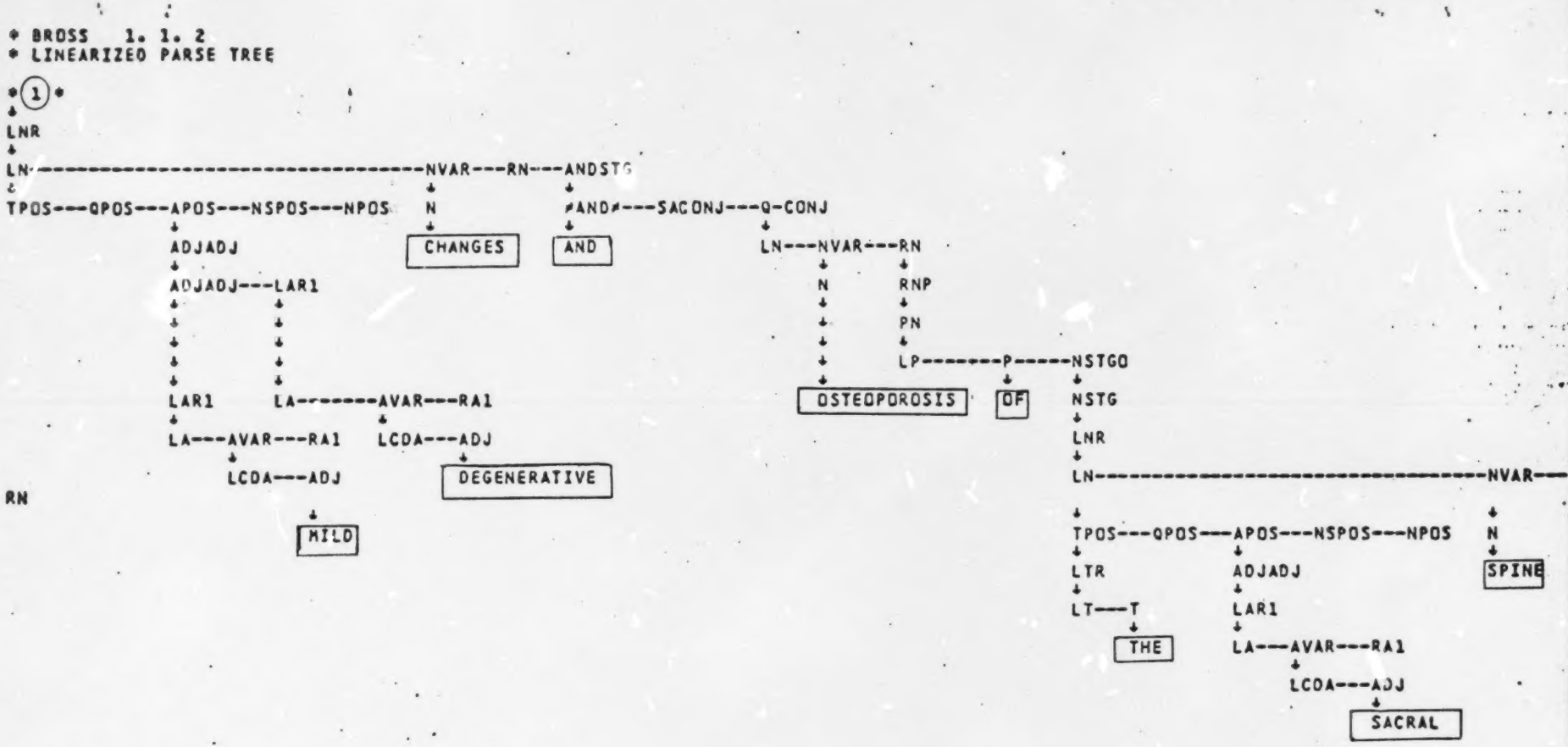

* BROSS 1. 1. 2
* LINEARIZED PARSE TREE
(1)
LNR
LN
TPOS
QPOS
APOS
NSPOS
NPOS
ADJADJ
ADJADJ
LAR1
LAR1
LA
AVAR
RA1
LA
AVAR
RA1
LCDA
ADJ
LCDA
ADJ
MILD
DEGENERATIVE
NVAR
RN
N
CHANGES
ANDSTG
ANDA
AND
SACONJ
Q-CONJ
LN
NVAR
RN
N
RNP
PN
LP
P
NSTGO
OSTEOPOROSIS
OF
NSTG
LNR
LN
NVAR
RN
TPOS
QPOS
APOS
NSPOS
NPOS
N
SPINE
LTR
ADJADJ
LT
T
LAR1
THE
LA
AVAR
RA1
LCDA
ADJ
SACRAL

Figure 4.5b

```
*BROSS 1.1.2  3-28-68 FLAT PLATE OF ABDOMEN -- MILD DEGERNATIVE CHANGES AND
*             OSTEOPOROSIS OF THE SACRAL SPINE .

FORMAT A
  ↓
DATA
  ↓
TEST ----------------------------------------FINDING
  ↓                                             ↓
TESTN---------TESTLOC----------TESTDATE    VERB------CHANGE-OVER-TIME
  ↓             ↓                ↓           ↓              ↓
[PLATE]--LADJ  [ABDOMEN]--LADJ  [3-28-6]    BE-SHOW      CHANGE
          ↓              ↓                    ↓              ↓
       [FLAT]         [OF THE]              [--]        [CHANGES]---LADJ
                                                                  ↓
                                                          [MILD DEGENFRATIVE]
```

```
FORMAT B
  ↓
DATA
  ↓
TEST ------------------------------------------FINDING
  ↓                                               ↓
TESTN----------TESTLOC----------TESTDATE    VERB------MED-FINDING----REGION
  ↓              ↓                ↓           ↓            ↓            ↓
[PLATE]---LADJ  [ABDOMEN]---LADJ  [3-28-6]    BE-SHOW    [OSTEOPOROSIS]  POSITION---PT-BODY
           ↓              ↓                    ↓                          ↓              ↓
        [FLAT]         [OF THE]              [--]                       [OF]         [SPINE]---LADJ
                                                                                            ↓
                                                                                      [THE SACRAL]
```

Figure 4.6

The format has two main parts, roughly corresponding to the subject and verb phrase of a full entry sentence, where the subject describes the TEST and the verb phrase describes the FINDING. Each part is further divided. In this example, the TEST part of the format has three filled slots: TESTNoun (flat plate), TESTLOCation (of the abdomen), and TESTDATE (3/28/68). Since the verb phrase contained a conjunction, the sentence was expanded into two sentences by the conjunction expansion transformation, resulting in two formats having the same TEST part but different FINDING parts. The first FINDING contains a CHANGE entry (mild degenerative changes); the second FINDING contains an entry in the MEDical FINDings slot (osteoporosis, recognized by the fact that osteoporosis is in a subfield word-class), and entries in REGION slots (of the sacral spine, also recognized by the occurrence of words in subfield word-classes).

This is not the place to discuss the formatting experiment in detail, since this is done elsewhere [44, 12]. But a few points are worth noting with regard to how this work differs from other uses of natural language processing. In clustering or indexing, for example, the correct assigning of a negative element does not affect the result, but here it is crucial. "No data" is very different from "No evidence of metastasis." The formatting therefore requires both a careful syntactic analysis and detailed subclassifications of the subfield words.

Of what use are such data structures? Properly organized within the computer (perhaps in more conventional data base

formalism), they would enable specific queries to be answered from the data, and statistical operations to be performed. In the health field such formatting might eventually be used to study the frequency and severity of recurring symptoms in certain diseases, as reflected in the patients' records, and to monitor on-going health care in clinics.

4.4 AIDING THE USER

Subscribers to retrieval services often have mixed reactions; they are glad to have the service because they need it and there is no alternative way to search the large literature store; at the same time they wish there was some way to "tune" the search more specifically to their needs. It has been suggested by many that what is needed is a facility whereby the user can direct the search himself. Similar remarks have been made about other uses of computers, such as tailoring or designing software packages for particular users. The naturally occurring image is that of the user before the console, interacting with the system and directing it (or being directed by it) by means of a dialogue carried out in natural language. This image is not universally applicable, however. Some users want no contact with computers; they just want better service. Both avenues are presently leading toward more emphasis on natural language processing.

Taking up first the matter of improving service, there is a distinct possibility that a natural language processor capable of handling scientific English could be used in

connection with existing retrieval systems to increase precision. Syntactic relations can be checked, and some retrieval errors, such as "false coordinations" might be caught this way. If this technique were perfected (and made more efficient), recall could be relaxed so that fewer relevant documents would be missed, and the processor would then be used to narrow down the set of retrieved documents to meet high precision standards. A caution here is a point made earlier: the problem of ambiguity in text processing remains. There is still the possibility that a great many selectional constraints are needed in order for parsing (or transformational analysis) to be feasible, or useful. And selectional constraints are precisely what is lacking in the language corpus of a diverse collection. On the other hand, queries and profiles can be sorted as to subject matter, and a "library" of selectional constraints slowly accumulated for the different subject areas. The view here is that the user could then be indirectly in natural language communication with the information sources, by virtue of a system which confronts the user's request with the natural language material of the retrieved documents. Limited experiments of this type could be set up using existing means.

One of the attractive aspects (at least at the outset) of the approach based on user interactive systems, is the dynamic features. Not only can the user see right away whether he is getting what he wants and re-direct the system if necessary,*

---

*As in the LEADERMART retrieval system developed by D. Hillman.

but even language ambiguity problems could be resolved by the user. In practice, the user is not going to want to look at sentence analysis trees or the equivalent, and answer questions about them--at least, not about very many. So a high quality language analyzer is a prerequisite for this avenue also. The interactive approach has a strong appeal because it allows the user to modify the request in response to what is available, but much remains to be understood about natural language interaction (discourse) before this solution will work well. There is interesting work in the area of artificial intelligence on discourse problems and language understanding. Some of this work may yet bear on practical problems of providing interactive "front ends" for information retrieval and information processing systems.

A few concluding remarks, on the important topic of processing times. Figure 4.7 shows a table of sample times for processing relatively simple sentence types. The top two-thirds of the table are a comparison made by Petrick [45] of published LSNLIS times with times we obtained for comparable sentences. Taking account of the differences in machine, time-sharing system, etc. (the IBM LISP system is faster) he concludes that a normalization would yield roughly comparable times for the two systems. The bottom section of the table shows some figures from LSP runs on sentences and sentence fragments of medium complexity. The times given in Fig. 4.7 should not be used to rate the efficiency of the different systems but to gain an

Figure 4.7

A few examples of Sentence Processing Times
(in seconds)

LSNLIS

Sentence	Parsing time	Interp. time	Total time
(1) How many lunar samples are there?	2.039	5.152	7.191
(2) How many breccias do not contain Europium?	6.272	8.593	14.865
(3) How many samples contain chromite?	3.579	8.277	11.856
(4) Which rocks contain chromite and ulvospinal?	7.743	9.782	17.525

REQUEST

Sentence			
(1) How many profitable companies are there?	3.868	0.163	4.031
(2) How many companies do not produce gas?	3.369	0.238	3.607
(3) How many companies sell computers?	3.226	0.239	3.465
(4) Which people sold IBM and Xerox?	3.046	0.189	3.235

LSP

Sentence or sentence fragment*	Parsing time	Format. time	Total time
(1) 5-2-67 chest--no change since 2-7-67.	1.19	2.89	4.08
(2) Chest X-ray 3-3-70 negative.	2.04	2.09	4.13
(3) 6-6 6-7-68 chest film 7-30 shows no improvement.	1.07	3.58	4.65
(4) 3-28-68 flat plate of the abdomen--mild degenerative changes and osteoporosis of the sacral spine.	6.52	8.38	14.90

*The numbers in these sentences are dates, and are analyzed as such by the system.

impression of the processing times that are currently possible. One sees that short sentences are processed in from 1 to 20 seconds. Long sentences from scientific articles, such as are parsed by the LSP, may run to over one minute. One should keep in mind, also, that research groups are using laboratory, not production, models.

If the times given here are representative of the current state of processing, the situation is not discouraging. Serious experiments in natural language information processing using existing processors are feasible, and it is not unreasonable to think of limited practical applications in the areas of question-answering from structured data bases, and automatic structuring of specialized natural language data bases. Mass processing of texts would not yet be economical if, indeed, it proves possible. However, we still have basic work to do on language processing, and in that time computer technology may catch up with our needs. To reverse a metaphor, one might say that if we can beat a path to an effective natural language analyzer, the world will build the better mousetrap (read ultra-fast special purpose computer) and lay it at our door.

5. RESEARCH GOALS IN LANGUAGE AND INFORMATION*

The discussion in this section begins with a reiteration of two obvious but important facts: Language is the major vehicle for storing and transmitting knowledge. It is also the natural means for interrogating an informational source. Thus, from the outset, language impinges on information delivery systems at its two end points: at the source end as the initial storage medium, in the form of documents; and at the user end as the natural form in which the user expresses an information need. To date, in most information systems, the language-information complex is left intact at both ends of the information channel, and identifying tags are substituted for the natural language texts and queries. While text-scanning has been shown to improve document retrieval and to make possible the retrieval of answer passages, so far the methods used do not "dissect" the language-information complex. Fundamental questions regarding the relation of information to language and to other symbolic forms become important as the goals of information technology advance, when we wish systems to be user-guided and to perform services of information extraction and repackaging (what has been called information regeneration) in addition to improved document retrieval.

---

*The material in this section was presented in a slightly different form at the Roundtable on Long-Range Research Planning in Information Science, sponsored by the NSF Division of Science Information, Anaheim, California, May 19-20, 1975.

It is my thesis that the new tasks of information systems can only be achieved on the basis of an increase in fundamental knowledge on how information is organized (or "coded") by language for storage, and how it is "decoded" by the reader or hearer, and combined with other knowledge to constitute useful information.

The relation of language structure to information is a central question. The question of whether there exist simpler forms than language (though probably language-like) for information, as suggested by recent research results, could be important for a future information-processing technology. A study of the different existing forms of information representation (especially where they are alternative forms for the same information) would be a stepping stone toward developing general forms for information representation, and useful in transforming and comparing information from various sources. The question of possible vocabulary factorization and reduction is another part of the general problem of regularizing the information representations, starting from the language form. All of these questions should be examined broadly, across the spectrum of the sciences, and also in detail, as they apply to particular subject matters.

The topics just mentioned are primarily concerned with the information-storage function of language and alternative forms for representing information. They can probably best be studied at the source end, i.e. in studies of textual material to answer such specific questions as where different types of information

are located in texts, what structural properties they have in the texts, and in what ways they can be combined. At the user end, where, in addition, dynamic aspects of language use come into play, a major question is how the different levels of language which are distinguished analytically (roughly: the syntactic, semantic and pragmatic) are integrated in an interactive situation. Here, semiotic studies and the use of interactive systems should be extended toward experimental situations involving real users and real data bases. At the syntactic-semantic bridge, one would like to know how much of the semantic burden is carried by the syntax common to all writing and speaking in a given language, how much by the particular constraints (syntactic and semantic) of scientific discourse, and how much by the specific semantics of the given subject matter. We also need to understand better how the syntactic and semantic components are integrated, and how semantic constraints can be entered into the system in a non-ad hoc manner. At the semantic-pragmatic bridge, the question arises as to how much outside information is implicitly drawn upon when users interact with the information presented to them. Also: what part of the additional information required in the decoding and matching process can be inferred from other information in the data base (inference systems) and what part, and what kind of information must be externally supplied? Practical questions as to how supplementary definitional and factual information can be made available to the system should also be investigated.

Advances in knowledge in the areas noted above will, I believe, lay the foundations for a new type of information technology. At the user end we can anticipate natural language interactive systems by means of which the user engages in a dialogue with the information store, initially to direct the literature search and then in a selected corpus of full texts to collate information from diverse documents according to categories and relations of special interest, some of which may evolve in the course of the dialogue. At the source end, we can envision that the documents (that is, titles, abstracts, or full-texts, depending on the task) will be capable of computer structuring into simpler language-like forms representing the major informational categories and connections in the material, suitable for search and processing. We can imagine a multi-level processor connecting the user and source ends, one that would be capable of sorting queries along several axes (type of search, form of answer, subject matter, level of detail), and ultimately be capable of aligning, comparing, and tabulating both linguistic and non-linguistic information, and reporting back to the user in the form of a summary table or in natural language.

Clearly, systems of the above type are beyond our present capabilities. However, they are not beyond reach if certain existing trends in technology and research are encouraged, and certain problem areas are anticipated and planned for. A few specific developments and problems can be mentioned.

First, at the source end, machine-readable natural language data bases are slowly coming into existence. As many

institutions convert their files to digital form for transactional efficiency, special purpose natural language data bases are thereby created. The existence of such data stores is in itself a pressure for more advanced processing techniques so that the data can be used. It is also likely that massive machine-readable natural language stores will be created as a byproduct of publishing operations. These could be saved and used in new systems equipped for natural language processing. However, there is a danger that this resource may not be fully exploited unless there is advance planning with regard to the economic, administrative, and social policy questions involved in their use. As techniques for processing natural language are developed, some attention should be paid to these policy questions.

In the area of forms of information representation, it has been fruitful to make a linguistic characterization of writing in the different sciences, and in different parts of articles within each science, and to compare the resulting structures with other information forms, such as tables, graphs, and partially structured material. Certain parts of the textual material are found to contain repetition of terms and term-classes in regular patterns which are characteristic of the particular subject matter, and are conveniently labeled "repeating information structures" (RIS). Roughly, as the material ranges from the "soft" to the "hard" sciences, and within an article from narrative and argument passages to result-presenting passages, the dominance of a few RIS patterns over other types of linguistic

material is evident. Also, the passages characterized by RIS's can often be aligned with tabular material presented in the same texts. This means (in gross terms) that while not all textual material in science and technology lends itself to natural language processing, those areas which do are capable of processing at a high level of informational specificity. It also indicates that for certain types of material there are likely to be underlying information structures into which both textual information and certain other forms of data (e.g. tables, charts, and perhaps graphs, but not diagrams or photographs) can be fitted.

In narrow subject matter areas (science subfields), the linguistic constraints are so marked that a grammar of specific usage can be stated for the subfield, in terms of which sentence formats with slots for different types of information are mechanically obtained for documents in the subfield. A combined linguistic-statistical program which has been developed can automatically generate semantic classes for such material. All of this considerably strengthens the view that information structures suitable for computer processing can be developed from the natural language form of information input. At the same time, one must not minimize the complexity of the language-reduction problem. Much detailed work of text analysis and proceduralization lies between the present promising results and operational systems based on the dissection of the language-information complex.

On the user end, the questions regarding how the different levels of language are integrated are in part suggested by

developments in artificial intelligence. While the AI systems to date operate in data areas which are indeed artificial from an information systems point of view, they are nevertheless breaking ground in the complicated area of interpreting user-generated natural language queries and commands. There is no doubt that the pressures toward natural language user-interfaces with computer systems will increase. In the information industry, the issue is not merely one of an English "front end," but one of supplying the user with a means to interact with the stored information. One may expect, therefore, that experiments with user-directed information systems and studies of the interactive process may profit from, and contribute to, studies of information structure based on automatic text analysis.

This discussion of research directions in semiotic and linguistic studies has proceeded mainly from the point of view of how these studies may contribute to new kinds of information systems in the future. However, I do not believe we will have to wait for the implementation of new systems in order to benefit from the results of such research. The problems of information representation and of user interaction with the system are already faced in existing systems. Deeper understanding of the language-information complex and of the query process may suggest improvements at both the source end (e.g. in the design of document descriptor languages) and the user end of document retrieval systems. Some new techniques which are developed in the course of the long range research may also be useful in adjunct operations, such as generating thesauri for specific

subject matters, and in automatic indexing. New information services, as techniques are developed, can be offered experimentally in special environments, both to test the new techniques and for the benefit of special groups of users.

At the same time that applicational efforts are encouraged, it should be clear that an important product of linguistic and NLP studies is the light they cast on scientific information. The "coding" of information in human language by scientists follows rules of linguistic structure and specialized word usage which have become part of the information itself. The language "code" is not an invented one, nor one designed to conceal rather than reveal its message. Rather it is found that the recording of information makes use of language structure as the general frame on which are hung the specific symbolic relations which carry the specific information. Subfield formats are a particular case in point. The study of this relation between form and content, between language structure and its semantic burden in scientific communication has already yielded tangible results. These results, in addition to spawning applications, are a clue that further study of this relation will contribute to a general characterization of the structure of scientific information.

REFERENCES

1. Panel: Can Present Methods for Library and Information Retrieval Service Survive?, Proceedings of the 1971 Annual Conference of the ACM, 564-567.

2. F.W. Lancaster, The Relevance of Linguistics to Information Science, Proceedings of the 1976 FID/LD Workshop on Linguistics and Information Science, in press.

3. Z.S. Harris, Mathematical Structures of Language, Section 5.6.2, Wiley Interscience, N.Y., N.Y., 1968.

4. G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill, New York, 1968.

5. K. Sparck Jones and M. Kay, Linguistics and Information Science, Academic Press, New York, 1973.

6. J. O'Connor, Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching, Information Processing & Management, vol. 11, 155-164, 1975.

7. R.F. Simmons, Answering English Questions by Computer: A Survey, Communications of the ACM, vol. 8, 53-70, 1965.

8. R.F. Simmons, Natural Language Question Answering Systems: 1969, Communications of the ACM, vol. 13, 15-30, 1970.

9. B. Henisz-Dostert and F.B. Thompson, The REL System and REL English, in A. Zampolli (ed.), Computational and Mathematical Linguistics, Proceedings of the International Conference on Computational Linguistics, Pisa, 1973, Casa Editrice Olschki, Florence, 1974.

10. S.R. Petrick, Semantic Interpretation in the REQUEST System, in A. Zampolli (ed.), Computational and Mathematical Linguistics, Proceedings of the International Conference on Computational Linguistics, Pisa, 1973, Casa Editrice Olschki, Florence, 1974.

11. W. Woods, R.M. Kaplan, and B. Nash-Webber, The Lunar Sciences Natural Language Information System: Final Report, Report #2378, Bolt Beranek and Newman, Cambridge, Mass., 1972.

12. L. Hirschman, R. Grishman, and N. Sager, From Text to Structured Information--Automatic Processing of Medical Reports, Proceedings of the National Computer Conference, AFIPS Press, Montvale, N.J., 1976.

13. N. Sager, Syntactic Formatting of Scientific Information. Proceedings of the 1972 Fall Joint Computer Conference, AFIPS Conference Proceedings vol. 41, 791-800, AFIPS Press, Montvale, N.J., 1972.

14. N. Sager, The Sublanguage Technique in Science Information Processing, Journal of the American Society for Information Science, vol. 26, 10-16, 1975.

15. Z.S. Harris, Discourse Analysis Reprints, Mouton, The Hague, 1963.

16. N. Sager, Perspective Paper on Computational Linguistics: Steps Towards Applications, Proceedings of the May 1976 FID/LD Workshop on Linguistics and Information Science, in press.

17. D.E. Walker, Automated Language Processing, Annual Review of Information Science, vol. 8, ed. C. Cuadra, American Society for Information Science, Washington, D.C., 1973.

18. F. J. Damereau, Automated Language Processing, Annual Review of Information Science, vol. 11, American Society for Information Science, Washington, D.C., 1976.

19. R. Grishman, A Survey of Syntactic Analysis Procedures for Natural Language, Courant Computer Science Report #8, Courant Institute of Mathematical Sciences, New York University, August, 1975.

20. T. Winograd, Understanding Natural Language, Academic Press, New York, 1972.

21. J.J. Robinson and S. Marks, PARSE, A System for Automatic Analysis of English Text, RM4564PR, Rand Corp. Santa Monica, Calif., 1965.

22. C. Raze, A Computational Treatment of Coordinate Conjunctions, American Journal of Computational Linguistics, to appear in 1976.

23. N. Sager, M. Salkoff, J. Morris, and C. Raze., Report on the String Analysis Programs, Introductory Volume, String Program Reports (S.P.R.) No. 1, Linguistic String Project, New York University and University of Pennsylvania, March, 1966.

24. N. Sager, Syntactic Analysis of Natural Language, Advances in Computers, vol. 8, 153-188, Academic Press, Inc., New York, 1967.

25. C. Raze, The FAP Program for String Decomposition of Scientific Texts, S.P.R. No. 2, Linguistic String Project, New York University, 1967.

26. W. Woods, An Experimental Parsing System for Transition Network Grammars, Natural Language Processing, ed. R. Rustin, Algorithmics Press, New York, 1973.

27. Z.S. Harris, Papers in Structural and Transformational Linguistics, D. Reidel, Dordrecht, 1970.

28. W.J. Plath, The REQUEST System, IBM RC 5604, IBM T.J. Watson Research Center, Yorktown Heights, N.Y., 1975.

29. R.F. Boyce et al., Specifying Queries as Relational Expressions, Proceedings of ACM SIGPLAN/SIGIR Interface Meeting, in SIGPLAN Notices, vol. 1, Jan. 1975, and in SIGIR Forum, vol. 9, winter 1974.

30. R.F. Boyce, D.D. Chamberlin, W.F. King III, and M.M. Hammer, Specifying Queries as Relational Expressions: The Square Data Sublanguage, Communications of the ACM, vol. 18, 621-628, Nov. 1975.

31. J. Hobbs, R. Grishman and C. Raze, The Automatic Transformational Analysis of English Sentences: An Implementation, International Journal of Computer Mathematics, in press.

32. L. Hirschman, R. Grishman and N. Sager, Grammatically-based Automatic Word Class Formation. Information Processing and Management, vol. 11, 39-57, 1975.

33. I.D.J. Bross and D.F. Stermole, Computer-Assisted Discourse Analysis of a Jargon, Computer Studies in the Humanities and Verbal Behavior, vol. IV, 2, 65-76, 1973.

34. N. Sager, Computerized Discovery of Semantic Word Classes in Scientific Fields. Directions in Artificial Intelligence: Natural Language Processing, Courant Computer Science Report No. 7, Courant Institute of Mathematical Sciences, New York University, 1975.

35. B. Bookchin, Computer Outputs for Sentence Decomposition of Scientific Texts, S.P.R. No. 3, Linguistic String Project, New York University, 1968.

36. N. Sager, A Computer String Grammar of English, S.P.R. No. 4, Linguistic String Project, New York University, 1968.

37. M. Salkoff and N. Sager, Grammatical Restrictions on the IPLV and FAP String Programs, S.P.R. No. 5, Linguistic String Project, New York University, 1969.

38. N. Sager, The String Parser for Scientific Literature, Natural Language Processing, R. Rustin, ed., Algorithmics Press, New York, 1973.

39. R. Grishman, The Implementation of the String Parser of English, Natural Language Processing, R. Rustin, ed., Algorithmics Press, New York, 1973.

40. R. Grishman, N. Sager, C. Raze, and B. Bookchin, The Linguistic String Parser, Proceedings of the 1973 Computer Conference, AFIPS Press, 1973, pp. 427-434.

41. E. Fitzpatrick and N. Sager, The Lexical Subclasses of the Linguistic String Parser, American Journal of Computational Linguistics; 2, 1974.

42. N. Sager and R. Grishman, The Restriction Language for Computer Grammars of Natural Language, Communications of the ACM, vol. 18, 390-400, 1975.

43. J. Chandioux, METEO: an Operational System for the Translation of Weather Forecasts, Proceedings of FBIS Seminar on the Present State of MT, 3/8-9/76, American Journal of Computational Linguistics, to appear.

44. B. Anderson, I.D.J. Bross, and N. Sager, Grammatical Compression in Notes and Records: Analysis and Computation, paper delivered at the 13th Annual Meeting of the Association of Computational Linguistics, Boston, Nov. 1, 1975, American Journal of Computational Linguistics, vol. 2, no. 4, 1975.

45. S.R. Petrick, On Natural Language Based Query Systems, IBM RC5577, Yorktown Heights, N.Y., 1975.